INTERNATIONAL SERIES OF MONOGRAPHS
ON PHYSICS • 87

Global Aspects in Gravitation and Cosmology

PANKAJ S. JOSHI

OXFORD SCIENCE PUBLICATIONS

Global Aspects in Gravitation and Cosmology

PANKAJ S. JOSHI
Tata Institute of Fundamental Research
Bombay, India

CLARENDON PRESS · OXFORD
1993

Oxford University Press, Walton Street, Oxford OX2 6DP
Oxford New York Toronto
Delhi Bombay Calcutta Madras Karachi
Kuala Lumpur Singapore Hong Kong Tokyo
Nairobi Dar es Salaam Cape Town
Melbourne Auckland Madrid
and associated companies in
Berlin Ibadan

Oxford is a trade mark of Oxford University Press

Published in the United States by
Oxford University Press Inc., New York

A catalogue record for this book is available from the British Library

Library of Congress Cataloging in Publication Data
Joshi, Pankaj S.
Global aspects in gravitation and cosmology / Joshi S. Pankaj.
(The International series of monographs on physics ; 87)
Includes bibliographical references and index.
1. Gravitation. 2. Cosmology. 3. Space and time. 4. General relativity (Physics). 5. Astrophysics. I. Title. II. Series: International series of monographs on physics (Oxford, England) ; 87.
QC178.J67 1993 531.1'4—dc20 93–37832
ISBN 0–19–853966–5

Typeset by the Author in TeX
Printed in Great Britain by
Biddles Ltd, Guildford and King's Lynn

PREFACE

The purpose of this book is to describe several basic results and applications of global aspects in gravitation theory and cosmology within the framework of Einstein's theory of gravity. Even though the discussion is based here mainly on general relativity, many of the results will hold for any metric theory of gravity based on a space-time manifold model.

The fundamental role played by global considerations in gravitation physics was clearly established by the theorems on space-time singularities developed by Hawking, Penrose, and Geroch, and the related theoretical advances towards understanding the structure of space-time. These developments are reviewed here in necessary detail to point out that major problems of significance remain, such as the nature and structure of space-time singularities, the cosmic censorship problem in black hole physics, and quantum effects in the very strong curvature fields near a singularity, an understanding of which is basic to any possible quantum gravity theory. Our treatment here of the issue of the final fate of gravitational collapse shows, by means of explicit consideration of several exact scenarios, that powerfully strong curvature naked singularities could result from the continual gravitational collapse of matter with several reasonable equations of state, such as describing the inflowing radiation, or dust or a perfect fluid. This places important constraints on possible formulations of the cosmic censorship hypothesis, and the final fate of gravitational collapse. This is a crucial open question in the general theory of relativity and relativistic astrophysics and is in fact basic to the validity of the theory and applications in black hole physics.

Several developments and new results on the structure and topology of space-time are reviewed here. Though the basic ingredients of this subject have stabilized now, we hope the treatment here will show that there is a wealth of new information to be gained yet. The classical theory of gravitation, namely the general relativity theory, admits singularities in space-time where the curvatures and densities could be infinite. The occurrence of singularities is generally considered to indicate an incompleteness in the theory and it is hoped that this problem may be solved in the full quantum gravity theory. Though such a theory is not available yet, we consider this alternative here by quantizing limited degrees of freedom of the metric tensor to examine the quantum effects near a singularity. Such quantum effects are typically shown to be diverging near a singularity for a wide range of space-times, giving rise to the possibility of singularity avoidance in quantum gravity. A separate chapter is devoted to working out applications of global techniques in cosmology, especially to obtain various model-independent upper limits on the age of the universe and to generate bounds on allowed particle mass values for the possible elementary particle clouds which may be invisibly filling the universe as the dark matter.

We hope that the treatment here will achieve two specific goals. Firstly, several important results on topological and causal structure of space-time, gravitational collapse and the cosmic censorship problem, global upperlimits in cosmology, etc. would have been reviewed and reported. This should create a live picture of global aspects in gravitation and cosmology and provide a fair idea of recent applications of global techniques in the general relativity theory. The second and equally important goal is to dispel an impression expressed sometimes that global techniques have been an area which was necessary only to prove singularity theorems and related results. In fact, throughout the book, an attitude will be maintained and a point of view emphasized that the very nature of gravitational force is such that global aspects of space-time inevitably come into the picture whenever we try to understand and interpret this force in detail. The point is, global aspects are inseparably interwoven with the very nature of gravitational force. Even in quantum gravity, the indications are such that global and non-perturbative considerations will be important there. While we have tried to make the treatment reasonably complete on a given topic, we have been somewhat selective in the choice of topics in view of the limitation on space. It is hoped, however, that some of the references would indicate other interesting directions in the global aspects in gravitation and cosmology.

The treatment here is intended to be self-contained and the basic definitions and concepts needed for the development of a specific topic are set up suitably at respective places. While the basic ingredients of general relativity are reviewed in Chapter 2, it is assumed that the reader has some elementary familiarity with the theory and some basic topological concepts. For further details on various aspects of general relativity and related topics we refer to Wald (1984), and to Weinberg (1972) and Narlikar (1983) for a detailed treatment on cosmology. The notations and terminology used here are that of Hawking and Ellis (1973), unless otherwise specified.

While writing this book I have benefited greatly at various stages from discussions with I.H. Dwivedi, Sonal Joshi-Desai, J.V. Narlikar, S.M. Chitre, P.C. Vaidya, Probir Roy, and several other friends. Many of the ideas reported have been worked out in collaboration with some of these colleagues and I thank them here. My understanding of the global aspects of gravity has benefited immensely from interactions with Robert Geroch and Chris Clarke. I recall with pleasure vigorous discussions with Ted

Newman and Jeff Winicour on issues concerning the asymptotic structure, and with G.M. Akolia, U.D. Vyas, and J.K. Rao on causal structure, of space-time.

It is my pleasure to thank several friends and colleagues at T.I.F.R. for their comments and encouragement, and in particular H. M. Antia, who made my interaction with computers so much easier. The need for a book of this nature was emphasized to me by J. V. Narlikar, who always considered global aspects to be important in gravitation and cosmology. I am grateful to him and P. C. Vaidya for their constant interest in the ideas expressed here and for their comments on the nature of gravity. I enjoyed working with Donald Degenhardt of the Oxford University Press, whose comments helped me to evolve a suitable plan for the book. I thank him and the OUP for excellent cooperation, and R. Preston for all the attention to the manuscript. The figures here were prepared by the T.I.F.R. drawing section.

P.S.J.

CONTENTS

1

GLOBAL THEMES IN GRAVITY

Global considerations have been important in theories of gravitation right from the inception of the general theory of relativity, and also in cosmology, by the very definition and purpose of that science. It was realized then that even though locally the laws of physics are those of special relativity and space-time is very nearly flat, the space-time universe as a whole is made by joining such local patches and gives rise to a non-flat, curved continuum which would also admit a suitable differential structure. The matter and energy density distribution is then described in terms of tensor fields on such a differentiable manifold. These matter fields in turn generate the space-time curvature via the field equations of Einstein. Similar global features will arise in other theories of gravity as well, such as those of Brans and Dicke (1961), which are metric theories of gravitation based on a space-time manifold model.

Thus, it turns out that in spite of being the weakest force of the known fundamental interactions (for instance, the ratio of the gravitational force and the electromagnetic force between an electron and a proton is 10^{-43}), gravity implies remarkable conclusions as far as the overall large-scale structure of the universe is concerned. For example, soon after the Einstein equations were discovered, Friedman (1922) showed that the universe must have originated a finite time ago from an epoch of infinite density and curvatures if the evolving matter obeys the dynamical equations of general relativity theory, together with the assumptions of homogeneity and isotropy.

In spite of such predominant global features evident in the structure of gravitation theory, most of the calculations were done, until the early 1960s, using a local coordinate system defined in the neighbourhood of a space-time event. Much of the effort was devoted then to solving Einstein equations using various simplifying assumptions, which form a rather complicated set of non-linear partial differential equations. The situation and approach changed considerably when the so-called 'Schwarzschild singularity' problem came up. The Schwarzschild exterior solution of Einstein's field equations describes the gravitational field outside a spherically symmetric star where there is no matter present and the space-time is empty. The space-time distance ds in (t, r, θ, ϕ) coordinates, between two infinitesimally separated events is given by the metric

$$ds^2 = -\left(1 - \frac{2m}{r}\right) dt^2 + \left(1 - \frac{2m}{r}\right)^{-1} dr^2 + r^2(d\theta^2 + \sin^2\theta\, d\phi^2). \quad (1.1)$$

Here m represents the mass of the star and the boundary of the star lies at $r = r_b$. The range of t and r coordinates is given by $-\infty < t < \infty$ and $r_b < r < \infty$, and θ and ϕ are the usual coordinates on the sphere. It is clear now that if $r_b < 2m$ or if eqn (1.1) represents the geometry outside that of a point particle of mass m placed at $r = 0$, then the above space-time has an apparent singularity at $r = 2m$ as seen by the divergence of metric component in eqn (1.1) at this value. It was thought initially that the above represents a singularity in the space-time itself and that physics goes seriously wrong at $r = 2m$. It was realized only after considerable effort that this is not a genuine space-time singularity but merely a coordinate defect, and what was really needed was an extension of the Schwarzschild manifold. This is indicated actually by the finiteness of curvature components at $r = 2m$. The point is, the coordinate system used above breaks down at this value of r and it describes only a patch of the space-time defined by the above coordinate range. Such an extension covering the rest of the space-time was obtained by Kruskal (1960) and Szekeres (1960) and this may be regarded as an important insight involving a global approach. Similar such developments which could be mentioned here are Alexandrov's (1950, 1967) work on space-time topology and the analysis of the Cauchy problem in general relativity (see for example, Wald (1984) for a review).

The study of global aspects in gravitation and cosmology really came into its own with a detailed analysis of the outstanding problem of space-time singularities forming in a space-time. As mentioned above, it was realized in the early days of relativity itself that an important implication of the study of cosmological models is that the universe contained an infinite curvature singularity from which it originated. The Schwarzschild solution mentioned above also contains a genuine curvature singularity at $r = 0$ where the space-time curvature components blow up as opposed to the coordinate defect at $r = 2m$. However, to begin with such singularities were not considered to be a serious physical problem in that they were thought to be a consequence of the exact symmetry conditions assumed while solving the Einstein equations, rather than being a genuine consequence of the general relativity theory. It was thought that these singularities would presumably disappear once more realistic conditions were used, replacing the exact symmetries such as the homogeneity and isotropy assumptions of the cosmological problem studied.

It was shown, however, by the work of Hawking, Penrose, and Geroch in the late 1960s and early 1970s, that this was not the case. By means of a rigorous analysis of global properties of a general space-time they showed that under certain very general and physically reasonable conditions such as the positivity of energy, occurrence of trapped surfaces, and a suitable causality condition, the space-time singularities must occur as an inevitable feature as far as a wide range of gravitation theories describing the gravitational force are concerned (Hawking and Ellis, 1973). Their analysis involved, on the one hand, a detailed examination of the causal structure of space-time, which was then combined with the study of gravitational focusing effect on the families of timelike and null geodesic congruences. The existence of space-time singularities then follows in the form of future or past incomplete non-spacelike geodesics in the space-time. Such a singularity would arise either in the cosmological scenario, where it provides the origin of the universe, or as the end state of the gravitational collapse of a massive star which has exhausted its nuclear fuel providing the pressure gradient against the inwards pull of gravity. Several global features for space-times were understood during this phase of work, which resulted in the development of singularity theorems, which have then been applied to further work on the evolution problem in relativity, black hole physics, structure of exact solutions of Einstein's equations, asymptotic structure of space-time, and such other areas.

While the above developments represent a notable advance towards understanding the global structure of space-time, outstanding problems of great significance remain unsolved as yet. In fact, apart from the existence of space-time singularities, we presently know very little about the global structure of Einstein's equations. The first and foremost amongst these problems is the question of nature and structure of space-time singularities as predicted by the singularity theorems. These theorems establish the existence of singularities in the form of either future or past incomplete non-spacelike geodesics for a very wide class of space-times. This would mean that the trajectories of freely falling material particles or photons will come to a sudden end which would disappear from the space-time. In the case of the material particles, this will happen after a finite amount of proper time as measured along their timelike trajectory. However, these theorems provide practically no information on the nature or the physical significance of the singularities predicted. For example, it is not known if the densities and curvatures will necessarily blow up along the non-spacelike trajectories falling into such singularities predicted by the singularity theorems. In fact, these quantities do diverge near the singularity for certain special classes of space-times which are physically relevant, such as for the Schwarzschild singularity at $r = 0$, and for many other families of exact solutions of Einstein's equations, as we will discuss in Chapters 3 and 5. Thus, the possibility cannot be ruled out that in many cases the singularity theorems would predict physically genuine singularities, though it is not clear at all under what conditions the theorems must give rise to such curvature singularities.

Another such fundamental question which yet remains unresolved is, even if such strong curvature singularities do occur in nature during the gravitational collapse of massive stars, whether they will always be necessarily hidden below the event horizons of gravity. The alternative is they could be naked and may be able to communicate with outside observers far away to affect the dynamics of the outside universe. In fact, in the Schwarzschild case discussed above, the singularity at $r = 0$, where the curvatures blow up, is completely covered by the event horizon forming at $r = 2m$ (see Fig. 1). Thus, no causal effects from the singularity can reach any outside observer for whom the singularity is totally hidden within a black hole which is invisible apart from its possible gravitational effects. This situation has special relevance to the gravitational collapse scenario, because, by Birkhoff's theorem, the geometry outside any spherically symmetric collapsing cloud must be necessarily Schwarzschild, described by the metric eqn (1.1). Thus, in the approximation of the star composed of homogeneous dust without any pressure, the curvature singularity forming as the end state of collapse will be completely covered by the event horizon and would be invisible to any external observer. The hypothesis that the above phenomenon is generically true, that is, the singularities forming in a general gravitational collapse should always be covered by the event horizons of gravity, and remain invisible to any external observer, is called the cosmic censorship hypothesis. This hypothesis, originally proposed by Penrose (1969), remains unproved as yet despite many attempts towards a proof, and has been recognized as one of the most important open problems in general relativity and gravitation physics. The point is, establishing the validity of this conjecture by means of a rigorous mathematical formulation and proof will confirm the already widely accepted and applied theory of black hole dynamics, which is finding considerable astrophysical applications presently. On the other hand, its overturn will throw the black hole dynamics into serious doubt.

Our objective here is to discuss the above and such other related important questions concerning the global aspects of gravity, and also to discuss some applications of the results which are already available on the global structure of space-time. In Chapter 2, we begin by setting up the space-time manifold model with a Lorentzian metric which is basic to all our considerations. Such a continuum structure of space-time has been verified to be valid up to the largest observable cosmological scales and on the smallest scale it would be valid at least until the quantum effects take over, which is much below the radius of an elementary particle of the order

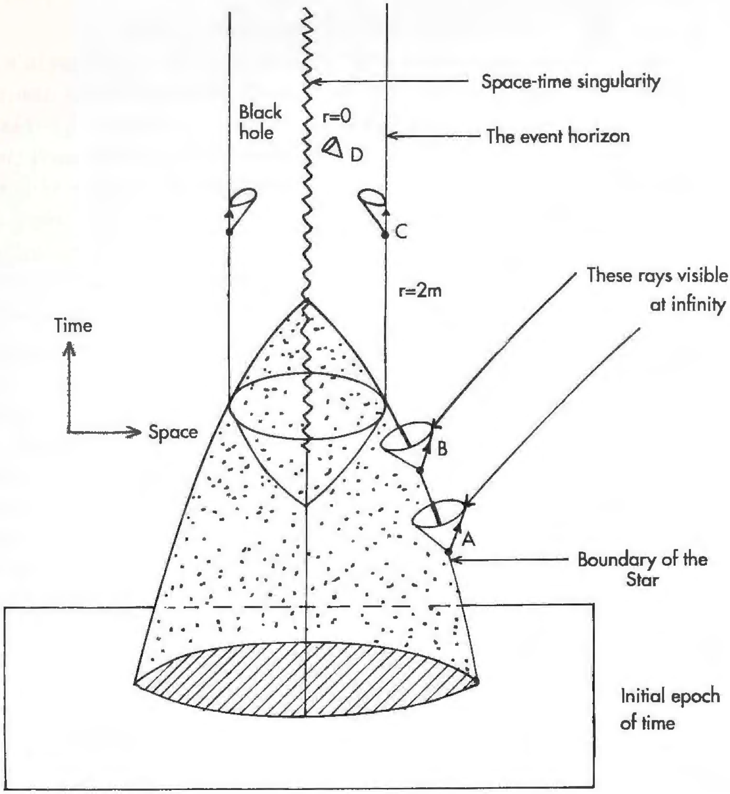

Fig. 1 Gravitational collapse of a massive star. The homogeneous dust cloud collapses from an initial epoch of time where the density at all points within the star and other physical quantities are all regular and well-behaved. The dotted region indicates the collapsing matter. Light rays emitted from points such as A and B on the surface of the star would reach a far away observer. However, as the collapse progresses, an absolute event horizon forms which is a wave front at $r = 2m$. No light ray from the point C can reach an outside observer and a black hole forms in the space-time, which is the empty region bounded by the event horizon at $r = 2m$. The space-time singularity forming at $r = 0$ due to the collapse of matter to infinite density is completely hidden below the event horizon and hence invisible to an outside observer. Both the ingoing and outgoing wave fronts from a point such as D will fall into the singularity.

of 10^{-13}cm. We introduce here necessary differential geometric concepts and results on manifolds which are needed later. Although we expect the reader to have a certain basic familiarity with the general relativity theory, the main ingredients of the theory are summarized here while developing the tools of a metric connection and the curvature tensor.

In Chapter 3, we examine several exact solutions of Einstein's equations for their properties, which will be useful later. This discussion also gives an idea of the possibilities present within the theory of relativity, such as the possible existence of space-time singularities and closed timelike curves. The discussion on the Minkowski space-time and the Schwarzschild space-time gives an idea of the asymptotic structure of these space-times, which is used later to define the general class of asymptotically flat space-times. In particular, for both these space-times, we work out the full families of null geodesics from an arbitrary apex in an explicit manner. Such families generate the future or past light cone from any given apex point which is a three-dimensional null hypersurface in the space-time.

The general theory of relativity implies the restriction that the space-time must be locally flat and the laws of special relativity must be valid in a coordinate patch, locally in a neighbourhood of any event. However, it places no restriction on the global topology of the space-time manifold. For the Minkowski space-time, which is the space-time of special relativity, the manifold is globally Euclidian with the topology of $\mathbb{R}^4$ and with the Lorentzian metric given by

$$ds^2 = -dt^2 + dx^2 + dy^2 + dz^2. \tag{1.2}$$

The causal structure at any event p is defined by the future and past light cones at p which are generated by the null geodesics at p given by the condition $ds^2 = 0$. All the events which p could influence by means of a timelike or null signal form the causal future of p, which is the interior of the future light cone at p, together with its boundary. Similarly, the causal past of p consists of all those events which could influence p with signals travelling at a speed less than or equal to that of light. In special relativity the future and past light cones of any event p never meet. However, if we allow for arbitrary topologies for the space-time, globally the future light cone of p may bend to enter the past of p, thus giving rise to causality violations. For example, consider the two-dimensional Minkowski space defined by the metric $ds^2 = -dt^2 + dx^2$ and where we identify the lines $t = -1$ and $t = 1$. This space is topologically $S^1 \times \mathbb{R}$ and contains closed timelike curves through every point (see Fig. 2). Such causality violations allow the observer to enter the past and often this is not considered physically desirable.

To avoid such causality violations and other related irregular behaviour, a number of regularity conditions are often placed on a space-time, which we study in Chapter 4. Such restrictions are often used as necessary conditions in the theorems which deduce the existence of space-time singularities (Hawking and Ellis, 1973; Beem and Ehrlich, 1981). Actually, it turns out

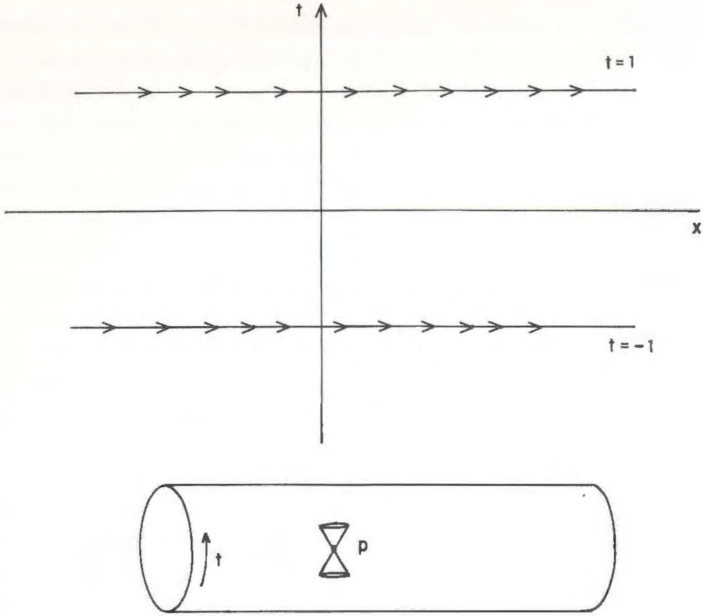

Fig. 2 A two-dimensional cylindrical space-time obtained from the Minkowski space-time by identifying $t = -1$ and $t = +1$. The arrows denote identification. For any point p, the future of p is the entire space-time and there are closed timelike curves through every event.

that not just one but a whole hierarchy of such conditions are required to ensure causal regularity of a space-time, and we look here for a unified condition which would ensure a globally well-behaved space-time. The properties of two physically interesting classes of models, namely, the reflecting and globally hyperbolic space-times, are studied here. The tool of causal functions on a space-time is developed and explored in detail here to show that this can be used to provide an overall unified characterization of the causal structure of space-time in terms of properties of these functions.

In Chapter 5, we review the developments on the occurrence of space-time singularities which include some generalizations of the conditions assumed originally to prove these theorems. The phenomena of causality violations is examined to show that under certain circumstances, the causality violating regions in a space-time will have a zero measure. A suitable causal regularity condition has always been one of the basic assumptions for the proofs of the singularity theorems. However, if any causality condition is violated only on a zero measure space-time set, such a violation need not be treated as being physically important. Further, it was shown by Tipler

(1977a), subject to certain special conditions, that when causality is violated, this itself causes space-time singularities in the form of null geodesic incompleteness. It may be noted, however, that some of the singularity theorems such as the Hawking theorem (1967), assume higher-order causality conditions such as the strong causality. We consider in this chapter the situations when such higher-order causality violations also would cause space-time singularities. Finally, we discuss here strong curvature conditions for singularities which characterize when a given space-time singularity should be considered physically important with significant physical consequences.

We address the issue of the final fate of gravitational collapse and the related cosmic censorship problem mentioned above in Chapters 6 and 7. A massive star starts gravitationally collapsing once it has exhausted its internal energy source of the nuclear fuel which generates the outwards pressure resisting the inwards gravitational pull of the matter. If the star could throw away most of its mass during the process of collapse, it could stabilize either to a white dwarf or neutron star configuration. However, for massive stars with tens of solar masses which exceed the Chandrasekhar limit of stability configuration, this appears quite unlikely because the star must throw away almost all its mass in such a process. Such stars must undergo a continual gravitational collapse to a space-time singularity of infinite curvature and density. Quantum gravity effects will become important in the very advanced stages of such a collapse at the scales of Planck length ($\sim 10^{-33}$cm). When these are taken into account, one could possibly avoid the eventual space-time singularity. However, a definite formulation of this idea is yet to be achieved which requires a suitable quantum gravity theory. In any case a region of matter as compact as 10^{-33} cm, where such quantum gravity effects start becoming important, may be considered to be a space-time singularity for all practical purposes. If such a singularity always occurs covered by the event horizon and hidden within a black hole, as it happens for the spherically symmetric homogeneous dust cloud collapse (see Fig. 1), the cosmic censorship and predictability in nature are preserved. On the other hand, occurrence of a naked singularity would imply a catastrophic breakdown of predictability in physics, because arbitrary bursts of radiation or matter could be radiated in the external universe by a naked singularity.

While black holes are already finding considerable astrophysical applications, as yet there is no general proof available for this censorship hypothesis which is at the very foundation of the physics of black hole. In fact, as we point out in Chapter 6, even a satisfactory formulation of this conjecture is not yet available and appears to be a formidable task to achieve. We conclude in this chapter that considerably more insight is needed in the phenomena of gravitational collapse before we are able to decide on the possible rigorous formulation and proof of the cosmic censorship hypothesis. Thus, we undertake a detailed examination of several gravitational collapse scenarios in this chapter, which include the collapse of radiation shells, general spherical self-similar collapse of a perfect fluid, and also non-self-similar collapse of inflowing radiation and inhomogeneous dust. We find a uniform pattern emerging from all these cases considered in that naked singularities do develop as the final end product of collapse. Further, in all the cases considered, not just isolated trajectories but families of non-spacelike geodesics come out of the naked singularity, providing a non-zero measure set of trajectories coming out. If only an isolated null trajectory came out of the naked singularity, that would amount to only a single wave front being emitted. On the other hand, the emission of a non-zero measure family of non-spacelike curves makes it a more serious phenomena.

An important test of the physical significance of such a naked singularity is its curvature strength. A detailed classification and analysis of space-time singularities is available as far as their curvature strength and physical significance are concerned (see for example, Clarke, 1975b; Tipler, Clarke and Ellis, 1980). A singularity is said to have the strong curvature property if there is at least one non-spacelike trajectory falling into it along which the space-time curvatures diverge sufficiently fast in the limit of approach to the singularity. We examine the strength of naked singularity in each of the above cases along all the non-spacelike trajectories terminating at the naked singularity in past, in the limit of approach to the singularity. It turns out that this is a strong curvature singularity in a very powerful sense in that curvatures diverge powerfully along not one but *all* non-spacelike geodesics meeting the naked singularity in past.

The above analysis leads to the remarkable conclusion that, in fact, strong curvature naked singularities can occur in general relativistic gravitational collapse for several reasonable equations of state. The scenarios considered here are reasonable in that the energy conditions are satisfied and the collapse evolves from a well-defined initial data. In fact, an interesting relationship emerges between the weak energy condition and the positivity of energy, and the occurrence of naked singularity with families of non-spacelike geodesics coming out of the singularity in past. Any possible formulation of the cosmic censorship hypothesis must take into consideration these situations and this leads us in Chapter 7 to examine the structure of a naked singularity and general constraints possible on the same. Thus, we examine there the relationship between topology change in a space-time and the occurrence of naked singularities and also the connection between the strength of naked singularities and the disruption in causal structure caused by the same. Variations of the censorship hypothesis in view of non-spherical collapse are also reviewed here.

We consider in Chapter 8 certain applications of global techniques in cosmology. General upper limits are obtained on important observable parameters such as the age of universe, and on masses of elementary particles which could fill the universe as dark matter, by using the results on the causal structure of space-time from Chapter 4. Traditionally, such limits are worked out in the framework of exact Friedmann models using the observed values of the Hubble constant H_0 and the deceleration parameter q_0. The observed values of these parameters are, however, subject to a great degree of uncertainty. Further, the recent observations of structures in the universe (see for example, Saunders *et al.* (1991), for a review) show that the universe could be inhomogeneous at the very large scales of even hundreds of mega parsecs. Even though it is conceivable that homogeneity is again recovered when the averaging is done on a still higher scale (statistical homogeneity), it would appear highly desirable to incorporate into the considerations the features such as perturbations from exact homogeneity. Towards this purpose, we model the observable universe by a general globally hyperbolic space-time where the spacelike surfaces of constant time need not admit the exact symmetries of homogeneity and isotropy. It is then shown that the use of only one observable parameter, such as the observed ages of globular clusters as obtained from nucleocosmochronology, imply quite meaningful and still fairly general cosmological upper limits such as those stated above.

In Chapter 9, we investigate the question of quantum effects near a space-time singularity. As yet, there is no complete quantum theory of gravity available. However, the quantum effects would be supposedly important in strong gravitational fields even before the full quantum gravity became operational at the Planck scales, and by examining the same we should get an idea of possible features present in full quantum gravity. Hence, we try here an approach quantizing only a limited degree of freedom of the metric tensor, namely the conformal factor, but examine the quantum effects near a singularity in a fairly general space-time scenario. Several exact situations are worked out and we examine the evolution of quantum effects in the black hole geometry as well. It is seen that a pattern emerges here which shows that in general the quantum effects considered would diverge in the vicinity of the space-time singularity. This offers the hope that even though the classical general relativity is beset with singularities, in quantum gravity there may be a possibility for singularity avoidance, which is examined here in some detail.

A note on the organization of the book is in order. As pointed out above, our main aim has been to report and review several useful developments involving different global aspects in gravity, which should create a live picture of research activity in this area. Thus, Chapter 9 can be read independently of earlier chapters, except that several concepts on global and causal structure from Chapter 4 and some exact models from Chapter 3 are used there. Similarly, Chapter 8 uses the results and concepts of Chapter 4, and the discussion on Robertson–Walker cosmological models from Chapter 3. Our treatment on the issue of the final fate of gravitational collapse is contained in Chapters 6 and 7. These chapters use to some extent the discussion on exact models from Chapter 3, and also some results on causality from Chapter 4, especially for Chapter 7, where we discuss general constraints on naked singularities. Readers familiar with general relativity and differential geometry might wish to return to Chapter 2 only when a familiarity with notation and terminology used is required.

2

THE MANIFOLD MODEL FOR SPACE-TIME

By the space-time manifold we mean a four-dimensional differentiable manifold M together with an indefinite Lorentzian metric tensor g which has signature $(-,+,+,+)$. The space-time (M,g) will be assumed to be space and time orientable and will obey several necessary topological regularity conditions such as the Hausdorffness, connectedness and so on, as we shall point out here. Such conditions provide physical reasonability to the space-time model being considered. In this chapter, we specify this basic model of space-time universe which is fundamental to the general theory of relativity and which underlies our considerations here on the global aspects in gravitation and cosmology.

The manifold model for space-time naturally incorporates the observed continuity of space and time and the essential principle of general relativity theory whereby the locally flat regions are glued together to obtain a globally curved continuum. In this case, a smooth change of coordinates is possible when a transition is made from one coordinate frame to the other. The existence of a Lorentz metric defines the causal relationships between events in the space-time by determining the past and future light cones for any event. While locally the geometry must always be that of the special theory of relativity, the global behaviour of light cones will depend on the choice of the metric tensor. Thus, even though locally the speed of light will not be exceeded at any given space-time event, globally the phenomena such as the occurrence of closed timelike curves and causality violation may be allowed in principle.

In Section 2.1 we introduce basic definitions concerning a differentiable manifold and vectors, and the topological and orientability properties of such a manifold are discussed in Section 2.2. After discussing tensors in Section 2.3, the basic role of metric tensor and the related metric connection are considered in Section 2.4. Timelike and null geodesics are a special set of non-spacelike trajectories which represent the motion of freely falling material particles and light rays and clarify many properties of a space-time. These are discussed in Section 2.5. Sections 2.6 and 2.7 discuss the topics of space-time diffeomorphisms and Killing vectors, and the curvature tensor. The Einstein equations are assumed to be satisfied on the space-time and this is discussed in Section 2.8.

2.1 Manifolds and vectors

The *n-dimensional Euclidian space* $\mathbb{R}^n$ is the collection of all n-tuples $(x^1, \ldots, x^n)$ such that $-\infty < x^i < \infty$, $i = 1, \ldots, n$; which is assumed to have the natural Euclidian metric. An *open ball* of radius r around any point x in $\mathbb{R}^n$ is the set of all points y such that $| x - y | < r$, where the modulus denotes the positive definite distance as defined by the Euclidian metric on $\mathbb{R}^n$. The *open sets* in $\mathbb{R}^n$ are sets which can be expressed as a union of such open balls.

An n-dimensional differentiable manifold is essentially a set which is locally similar to an open set of $\mathbb{R}^n$. That is, locally Euclidian patches are glued together smoothly to obtain a space which need not be Euclidian globally. Specifically, an *n-dimensional, C^∞, real differentiable manifold* is a set M together with a collection $\{u_\alpha, \phi_\alpha\}$, (called an *atlas* for M, where u_αs are subsets of M and ψ_α is a one-one map of a given u_α onto an open subset in $\mathbb{R}^n$) such that:

1. the sets u_α form a cover for M, that is any given p in M lies in a u_α for some α,
$$M = \bigcup_\alpha u_\alpha; \tag{2.1}$$
2. whenever u_α and u_β intersect, that is, $u_\alpha \cap u_\beta \neq \emptyset$, then the map $\phi_\alpha \circ \phi_\beta^{-1}$ from $\mathbb{R}^n$ to $\mathbb{R}^n$, which takes points of $\phi_\beta(u_\alpha \cap u_\beta)$ to points of $\phi_\alpha(u_\alpha \cap u_\beta)$ is infinitely differentiable in a continuous manner (that is, it is a smooth function, which is also denoted as a C^∞ function) as a mapping between two open subsets of $\mathbb{R}^n$ (see Fig. 3).

One could have considered alternatively the map $\phi_\beta \circ \phi_\alpha^{-1}$ and again the same condition would hold. Each u_α is called a *local coordinate neighbourhood* or a *chart* where $p \in u_\alpha$ has coordinates of $\phi_\alpha(p)$ in $\mathbb{R}^n$. The condition (2) above ensures that whenever $p \in M$ undergoes a coordinate change, it must be smooth. Alternatively, if $\{x^i\}$ and $\{y^i\}$ are local coordinates of $p \in M$ in u_α and u_β respectively, then the functions $x^i = x^i(y^1, \ldots, y^n)$ are C^∞ functions from $\mathbb{R}^n$ to $\mathbb{R}^n$. It is customary to choose a *maximal* or *complete atlas* for a given manifold M; that is, if $\{u_\alpha, \phi_\alpha\}$ is an atlas for M, one chooses for M the atlas which consists of all other atlases compatible with $\{u_\alpha, \phi_\alpha\}$ (that is, their union with $\{u_\alpha, \phi_\alpha\}$ is also a C^∞ atlas). Such a choice ensures that one has included all possible, mutually compatible, coordinate systems for a given manifold M. A C^r-manifold is defined in a similar way by requiring that the transition functions $\phi_\alpha \circ \phi_\beta^{-1}$ are r-times continuously differentiable (a continuous function is denoted by C^o).

The plane $\mathbb{R}^2$, or the Euclidian space $\mathbb{R}^n$ is of course a manifold in its own right as it is covered by a single chart $\mathbb{R}^n$ itself, where ϕ would be the identity map with the coordinate range being $-\infty < x^i < \infty$ for

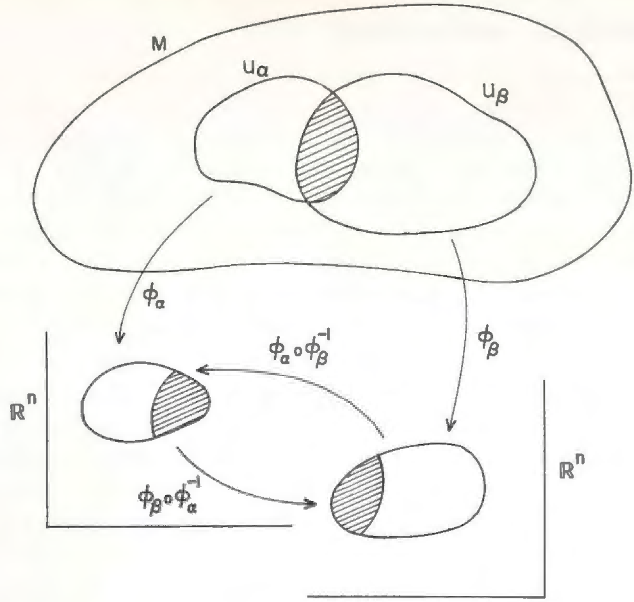

Fig. 3 The smooth maps $\phi_\alpha \circ \phi_\beta^{-1}$ on the n-dimensional Euclidian space $\mathbb{R}^n$ giving the change of coordinates in the overlap region.

$i = 1, \ldots, n$. As an illustrative example of the above ideas, consider the 2-sphere S^2 defined by

$$S^2 = \{(x^1, x^2, x^3) \in \mathbb{R}^3 \mid (x^1)^2 + (x^2)^2 + (x^3)^2 = 1\}. \tag{2.2}$$

The six hemispherical open sets $O_i^\pm$ for $i = 1, 2, 3$ are given by $O_i^\pm = \{(x^1, x^2, x^3) \in S^2 \mid \pm x^i > 0\}$ which cover S^2. Each $O_i^\pm$ is mapped as required onto the open disc $\{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 < 1\}$ by the projection maps given by $f_1^+(x^1, x^2, x^3) = (x^2, x^3)$ etc. The overlap functions $f_i^\pm \circ (f_j^\pm)^{-1}$ are seen to be C^∞ functions in their domain of definition. Thus, S^2 is a two-dimensional, C^∞ manifold which cannot be covered by a single coordinate system. Similarly, the n-dimensional sphere S^n is also seen to be a differentiable manifold.

We give a few more definitions. A function $f : M \to \mathbb{R}$ is called *differentiable* if the map $f \circ \phi_\alpha^{-1}$ is a C^∞ map for all charts ϕ_α as a map from $\mathbb{R}^n$ to $\mathbb{R}$. One can similarly define C^r functions (Spivak, 1965). Next, suppose, M and M' are two differentiable manifolds with ϕ_α and ψ_α denoting charts of M and M' respectively. A map $h : M \to M'$ is called a *C^r-differentiable map* if the map $\psi_\alpha \circ h \circ \phi_\alpha^{-1}$ is always C^r differentiable as a map from $\mathbb{R}^n$ to $\mathbb{R}^n$ for all α. If the dimension of M is n and that of M' is n' with $n > n'$,

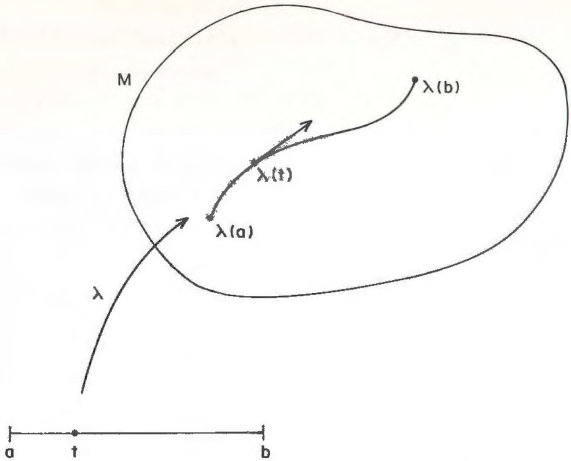

Fig. 4 A curve in a differentiable manifold.

then the map h cannot be one-one. However, if h is a one-one, onto and continuous map from M to M' such that h^{-1} is also continuous, then h is called a *homeomorphism.* If a homeomorphism h and its inverse h^{-1} are both C^r maps, then h is called a C^r-*diffeomorphism.*

A C^k-*curve* in M is a C^k map from an interval of $\mathbb{R}$ into M (Fig. 4). A *vector* (or a *contravariant vector*) $(\partial/\partial t)_{\lambda(t_0)}$ which is tangent to a C^1-curve $\lambda(t)$ at a point $\lambda(t_0)$ is an operator from the space of all smooth functions on M into $\mathbb{R}$ given by

$$\left(\frac{\partial}{\partial t}\right)_{\lambda(t_0)}(f) = \left(\frac{\partial f}{\partial t}\right)_{\lambda(t_0)} = \lim_{s\to 0}\frac{f[\lambda(t+s)] - f[\lambda(t)]}{s}. \tag{2.3}$$

This is basically $\frac{d}{dt}(f \circ \lambda)$, which is the derivative of f in the direction of $\lambda(t)$ with respect to the parameter t. Choosing $f = t$, where t is the parameter along the curve, we have

$$\left(\frac{\partial}{\partial t}\right)_{\lambda}(t) = 1.$$

If $\{x^i\}$ are local coordinates in a neighbourhood of $p = \lambda(t_0)$, then

$$\left(\frac{\partial f}{\partial t}\right)_{\lambda(t_0)} = \frac{dx^i}{dt}\frac{\partial f}{\partial x^i}\Big|_{\lambda(t_0)}, \tag{2.4}$$

where a repeated index implies summation over the values $1, \ldots, n$. (We shall use this *summation convention* throughout.) Thus, every tangent vector at $p \in M$ can be expressed as a linear combination of the coordinate derivatives, $(\partial/\partial x^1)_p, \ldots, (\partial/\partial x^n)_p$. Conversely, one could choose any linear combination of these operators which are partial derivatives with respect to coordinates, namely, $V^i(\partial/\partial x^i)_p$, where V^i are any numbers. It is then possible to find a curve which admits this linear combination as a tangent (see for example, Wald (1984) for details). Further, the vectors $(\partial/\partial x^j)_p$ are linearly independent; if this is not the case then there are numbers V^i such that

$$V^i \left(\frac{\partial}{\partial x^i}\right)_p = 0, \tag{2.5}$$

with at least one V^i non-zero. Then, applying this to the coordinate functions, $x^1, \ldots, x^n$, we get $V^k = 0$ for all k, which is a contradiction. Thus, the vectors $(\partial/\partial x^j)$ span the vector space T_p, which is the space of all tangent vectors at p. Then the vector space structure is defined by

$$(\alpha X + \beta Y)f = \alpha(Xf) + \beta(Yf). \tag{2.6}$$

The vector space T_p is also called the *tangent space* at p. The basis $\{(\partial/\partial x^i)_p\}$ is called a *coordinate basis* of T_p, whereas we denote a general basis by $\{e_i\}$, where $i = 1, \ldots, n$, are linearly independent vectors. Then, for any vector $V \in T_p$,

$$V = V^i e_i, \tag{2.7}$$

where the quantities V^i are called the components of V with respect to the basis e_i. In a coordinate basis, $V^i = dx^i/dt$. Again, $\{\partial/\partial x^i\}$ forms a basis of T_p which implies that the dimension of T_p is n.

Given the tangent space T_p at $p \in M$, one can define naturally the vector space of all the dual vectors at p, also called *covariant vectors* or *one-forms* at p. A one-form ω at p is a real-valued linear function on T_p, denoted by $\omega(X) \equiv \langle \omega, X \rangle$, and the linearity condition implies,

$$\langle \omega, \alpha X + \beta Y \rangle = \alpha \langle \omega, X \rangle + \beta \langle \omega, Y \rangle, \tag{2.8}$$

for $\alpha, \beta \in \mathbb{R}$ and $X, Y \in T_p$. Given a tangent space basis $\{e_a\}$, a unique set of one-forms $\{e^a\}$ is defined by the condition that the given one-form e^b maps a vector V into V^b, that is, the bth component of V in the basis e_a. Thus

$$\langle e^b, V \rangle = V^b. \tag{2.9}$$

From the above we get

$$\langle e^a, e_b \rangle = \delta^a{}_b.$$

The linear combinations of one-forms are defined by

$$\langle \alpha\omega + \beta\eta, V\rangle = \alpha\langle \omega, V\rangle + \beta\langle \eta, V\rangle, \tag{2.10}$$

with $\alpha, \beta \in \mathbb{R}$. Then $\{e^a\}$ may be regarded as a basis of the space of all one-forms at p because one could write any one-form ω as $\omega = \omega_a e^a$ with

$$\omega_a = \langle \omega, e_a\rangle.$$

Thus, the set of all one-forms at the event p forms a vector space at p which is *dual* of T_p and we denote it by T_p^*. The basis e^a is called a *dual basis* to e_a. If $\omega \in T_p^*$ and $V \in T_p$, we have

$$\langle \omega, V\rangle = \langle \omega_a e^a, V^b e_b\rangle = \omega_a V^b \delta^a{}_b = \omega_a V^a. \tag{2.11}$$

A *vector field* V on a manifold M is an assignment of a tangent vector V_p at each $p \in M$. The vector field is said to assign vectors smoothly if for each smooth function f on M, the function $V(f)$ is also smooth on M, which is the directional derivative of f along the vector V_p at each point p. Now, the coordinate basis vector fields $\partial/\partial x^i$ are smooth, and hence a vector field will be smooth provided its coordinate components V^i are smooth functions. Given two vector fields V and W, a new vector field, called their *commutator* $[V, W]$, is defined by

$$[V, W](f) = V[W(f)] - W[V(f)]. \tag{2.12}$$

We note that the commutator for any two coordinate basis vector fields vanishes. If f and g are any two smooth functions, it is easy to see that $[V, W](f + g) = [V, W](f) + [V, W](g)$ and also that $[V, W](\alpha f) = \alpha[V, W](f)$ for any $\alpha \in \mathbb{R}$. One could also show that

$$[V, W](fg) = f[V, W](g) + g[V, W](f),$$

which is the product property. It can be seen by expanding in a coordinate basis that the commutator $[V, W]$ will be a smooth vector field if and only if both V and W are smooth. We also note that $[V, V] = 0$ and that $[V, W] = -[W, V]$. Further, the commutator is linear in each of its arguments with respect to addition and we have

$$[V_1 + V_2, W] = [V_1, W] + [V_2, W].$$

Any smooth function f on M defines a one-form df, called the *differential* of f, by the rule

$$\langle df, V\rangle \equiv Vf. \tag{2.13}$$

Thus, in a coordinate basis we have

$$\langle df, V\rangle = V^a \frac{\partial f}{\partial x^a}. \tag{2.14}$$

The local coordinate functions $(x^1, \ldots, x^n)$ can be used to define a set of one-forms $(dx^1, \ldots, dx^n)$, which yields a basis dual to the coordinate basis. This is because

$$\langle dx^a, \frac{\partial}{\partial x^b}\rangle = \frac{\partial x^a}{\partial x^b} = \delta^a{}_b. \tag{2.15}$$

Also, this gives

$$df = \langle df, \frac{\partial}{\partial x^a}\rangle dx^a = \frac{\partial f}{\partial x^a} dx^a, \tag{2.16}$$

which is the usual definition of the differential df. If f is a non-constant function, then the surfaces $f =$ const define an $(n-1)$ dimensional submanifold of M. Consider now the set of all the vectors $V \in T_p$ such that

$$\langle df, V\rangle = Vf = 0,$$

then, the vectors V are tangent to curves in the $f =$ const. submanifold, through p. In this sense the differential df is normal to the surface $f =$ const. at p.

2.2 Topology and orientability

Given a C^∞ maximal atlas on a space-time manifold M, there is a natural *topology on M* defined by the companion Euclidian space by requiring each ϕ_α to be a homeomorphism. Thus, the open sets in M are pre-images of open sets in $\mathbb{R}^n$, and their unions. This is equivalent to the statement that the collection $\{u_\alpha\}$ provides a basis for topology of M. Here $\mathbb{R}^n$ is taken to have its canonical topology defined by the metric

$$d(x, y) = [(x_1 - y_1)^2 + \ldots + (x_n - y_n)^2]^{1/2}, \tag{2.17}$$

for any $x, y \in \mathbb{R}^n$.

In the following, we specify several topological regularity conditions which we assume for a physically reasonable space-time manifold. The space-time will be assumed to be *Hausdorff*, that is, given p and q with $p \neq q$ in M, there are disjoint open sets u_α and u_β in M such that $p \in u_\alpha$ and $q \in u_\beta$. Physically interesting known space-time examples such as the Schwarzschild geometry and Robertson–Walker space-times (to be discussed in Chapter 3) are topologically Hausdorff. This is a reasonable requirement on a space-time which ensures the uniqueness of limits of convergent sequences and incorporates our intuitive notion of distinct space-time events.

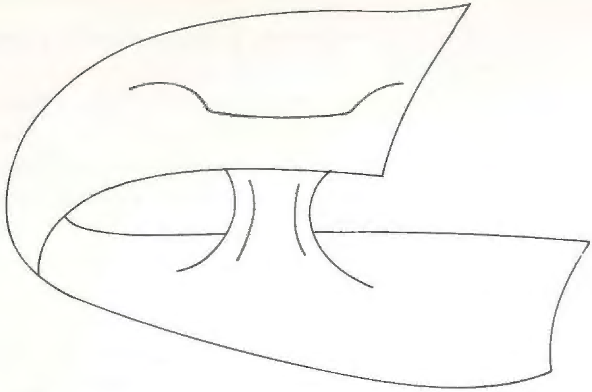

Fig. 5 A two-dimensional wormhole in the topology of space.

Next, the space-time manifold M has *no boundary*. This is reasonable because a boundary represents in a sense the 'edge' of the universe not detected by any astronomical observations. It is common to have manifolds without boundary, for example, for a two-sphere S^2 in $\mathbb{R}^3$, no point on S^2 is a boundary point in the induced topology on the same as implied by the natural topology on $\mathbb{R}^3$. All the neighbourhoods of any $p \in S^2$ will be contained within S^2 in this induced topology. We also assume M to be *connected*, that is, one cannot have $M = X_1 \cup X_2$, with X_1 and X_2 being two open sets and $X_1 \cap X_2 = \phi$. This is because disconnected components of the universe cannot interact by means of any signals and the observations are confined to the connected component wherein the observer is situated. It is not known, however, if M could be simply-connected or multiply-connected. In fact, Wheeler (1962) and Misner, Thorne and Wheeler (1973) considered multiply-connected space-times in detail and introduced the notion of a wormhole in the Schwarzschild geometry. Such wormholes are like 'handles' in the multiply connected topology of space and could connect widely separated regions in space (Fig. 5).

It is known, however, that such wormholes will not be stable and will collapse as soon as they are created unless one allows for the violation of the energy condition in an averaged sense, which would imply the existence of negative energy fields (Morris, Thorne and Yurtsever, 1988; Deutsch and Candelas, 1980; Lee, 1983). Thus, one could probably stabilize such a wormhole only by shifting the energy of vacuum to be negative by using quantum processes. Another possibility which could give rise to a multiply-connected space-time is the process of topology change which we consider in some detail in Section 7.3. It is not clear again if the topology of space could change while it evolves in time, and if so, what physical agencies could cause such a change. However, as we point out in Chapter 7, if topology change is allowed, it affects the structure of space-time in a severe manner in the sense of causing naked singularities.

A space-time will always be assumed to be *non-compact.* This is because compact space-times violate causality and admit closed timelike curves, as will be shown in Chapter 4. Such an observer can enter his or her own past, which is considered to be highly unphysical. It is also usual to assume that M is paracompact. An atlas $\{u_\alpha, \phi_\alpha\}$ is called *locally finite* if there is an open set containing every $p \in M$ which intersects only a finite number of the sets u_α. A manifold M is called *paracompact* if for every atlas $\{u_\alpha, \phi_\alpha\}$, there is a locally finite atlas $\{O_\beta, \psi_\beta\}$ with each O_β contained in some u_α For a further discussion on these and related topological concepts, we refer to Simmons (1963) or Willard (1970). For a connected, Hausdorff manifold, the paracompactness condition is actually equivalent to the existence of a countable base for the topology of M. One would like to consider this as a reasonable assumption on M because we assume (Section 2.4) the existence of a Lorentz metric tensor defined globally on M, and any Hausdorff manifold with a C^r Lorentz metric can be shown to be paracompact (Geroch, 1968b).

Let B be the set of all ordered basis $\{e_i\}$ for T_p. If $\{e_i\}$ and $\{e_{j'}\}$ are in B, then we have

$$e_{j'} = a^i_{j'} e_i. \tag{2.18}$$

If a denotes the matrix $[a^i_j]$, then $\det[a] \neq 0$. We introduce an equivalence relation in B by the condition that $e_i \sim e_{j'}$ if and only if $\det[a] > 0$. Clearly, there are exactly two such equivalence classes which are called the *orientations* of T_p. By an arbitrary choice, one of these classes is called a *positive orientation* and the other one is called a *negative orientation* for T_p. Now, let M be a manifold and $p \in M$. If for all p there is a neighbourhood U of p and n continuous linearly independent vector fields $\{\xi_1, \ldots, \xi_n\}$ such that for any $q \in U$, the basis $\{\xi_1(q), \ldots, \xi_n(q)\}$ belongs to the same equivalence class under a transformation of coordinates when one chooses another chart V containing q. In such a case, under all possible coordinate transformations, the determinant for the transformation matrix of the basis will have the same sign. In this case, M is called an *orientable* manifold. This definition could then be stated in the following equivalent form. An n-dimensional manifold M is called orientable if M admits an atlas $\{U_i, \phi_i\}$ such that, whenever $U_i \cap U_j \neq \emptyset$, then

$$J = \det\left(\frac{\partial x^i}{\partial x^j}\right) > 0, \tag{2.19}$$

where $\{x^i\}$ and $\{x^j\}$ are local coordinates in U_i and U_j respectively.

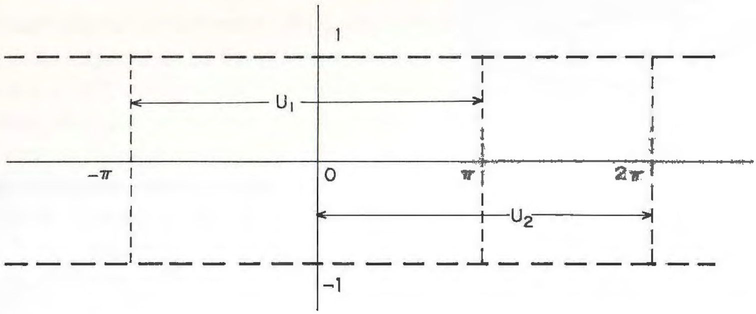

Fig. 6 The Möbius strip is a non-orientable manifold. A basis of vectors at any point does not have the same orientation when it comes back after going round the strip.

The Möbius strip is an example of a non-orientable manifold (Fig. 6). Take $Z = \{(x, y) \in \mathbb{R}^2 \mid -1 < y < 1\}$ in the two-dimensional Euclidian space and introduce the following identification: $(x, y) \sim (x+2\pi, -y)$. Take $M = Z/\sim$ and introduce the charts

$$U_1 = \{(x, y) \mid -\pi < x < \pi\}, \tag{2.20}$$

$$U_2 = \{(x, y) \mid 0 < x < 2\pi\}. \tag{2.21}$$

Then $U_1 \cup U_2$ covers M and the maps $\phi_1 : U_1 \to (-\pi, \pi) \times (-1, 1)$ and $\phi_2 : U_2 \to (0, 2\pi) \times (-1, 1)$ are homeomorphisms which define an atlas for M. The region $U_1 \cap U_2$ is the strip $(0, 2\pi) \times (-1, 1)$ except for the line $x = \pi$. On the first region of intersection, we have $(\phi_1 \circ \phi_2^{-1})(x, y) = (x, y)$ and the transformation matrix has determinant $+1$. On the second region, $(\phi_1 \circ \phi_2^{-1})(x, y) = (x - 2\pi, -y)$ and the Jacobian matrix has determinant -1. In general, one could see that a vector defined at a point on the strip with a positive orientation comes back with a reversed orientation in negative direction when it traverses along the strip to come back to the same point.

2.3 Tensors

Tensors are geometric objects defined on a manifold, which remain invariant under the change of coordinates. There are various matter fields defined on a space-time such as the electromagnetic fields or dust and so on, which are represented by the stress-energy tensor of the space-time. On the other hand, the global geometry and curvature of the manifold are described by fields such as the metric tensor (which we discuss in the next section), and the curvature tensor. In the general theory of relativity it is required that the form of physical laws must remain unchanged under a general transformation of coordinates (principle of general covariance). Thus, physical fields are represented by various tensor fields on the space-time and the laws governing them are expressed as tensor equations which remain valid under arbitrary coordinate transformations. When one specializes to an inertial coordinate system, these laws reduce to the equations of special relativity.

In general, a *tensor* T of (r,s) type at $p \in M$ is a multilinear real-valued function on the Cartesian product

$$T_p^* \times \ldots \times T_p^* \times T_p \times \ldots \times T_p \rightarrow \mathbb{R}, \tag{2.22}$$

where there are r-factors of T_p^* and s-factors of T_p. Thus, T acts on one-forms and vectors in general to produce a real number.

If T is a tensor of (r,s) type at $p \in M$,

$$T(\omega_1,\ldots,\omega_r,V_1,\ldots,V_s) = T(\omega_{i_1}e^{i_1},\ldots,\omega_{i_r}e^{i_r},V^{j_1}e_{j_1},\ldots,V^{j_s}e_{j_s}). \tag{2.23}$$

Using the multilinear property of T, the above can be written as

$$T^{i_1\ldots i_r}{}_{j_1\ldots j_s}\omega_{i_1}\ldots\omega_{i_r}V^{j_1}\ldots V^{j_s},$$

where we have defined

$$T^{i_1\ldots i_r}{}_{j_1\ldots j_s} \equiv T(e^{i_1},\ldots,e^{i_r},e_{j_1},\ldots,e_{j_s}), \tag{2.24}$$

and $\{e_i\}$ and $\{e^i\}$ are basis vectors at p for T_p and T_p^* respectively.

The space of all the tensors of type (r,s) at p is called the *tensor product* $T_s^r(p)$ and is denoted by

$$T_s^r(p) = T_p \otimes \ldots \otimes T_p \otimes T_p^* \otimes \ldots \otimes T_p^*, \tag{2.25}$$

where there are r-factors of T_p and s-factors of T_p^*. The dimension of T_s^r is n^{r+s}, where n is the dimension of the manifold. This is a vector space of all (r,s) tensors over real numbers with the addition of tensors and scalar multiplication defined in a natural manner. In particular, a vector is a tensor of type (1,0) and a one-form is a tensor of types (0,1). In terms of the basis vectors $\{e_i\}$ and $\{e^i\}$ for the tangent space and cotangent space at p, the set

$$\{e_{i_1}\otimes\ldots\otimes e_{i_r}\otimes e^{j_1}\otimes\ldots\otimes e^{j_s}\},$$

forms a basis for the tensor product $T_s^r(p)$ with all the indices running from 1 to n. Then, any tensor $T \in T_s^r$ can be expressed as

$$T = T^{i_1\ldots i_r}{}_{j_1\ldots j_s}e_{i_1}\otimes\ldots\otimes e_{i_r}\otimes e^{j_1}\otimes\ldots\otimes e^{j_s},$$

where the tensor components $T^{i_1 \dots i_r}{}_{j_1 \dots j_s}$ are defined as above.

Consider now a change of coordinates, which causes a change of basis $\{e^i\}$ to $\{e^{i'}\}$ and similarly, $\{e_j\}$ going to $\{e_{j'}\}$. In particular, let us choose a coordinate basis $\{\partial/\partial x^i\}$ for T_p and corresponding basis $\{dx^i\}$ for the cotangent space T_p^*. Then, under a change of coordinates, the components of T in the new coordinates $\{x^{i'}\}$ can be written as

$$T^{i'_1 \dots i'_r}{}_{j'_1 \dots j'_s} = T\left(dx^{i'_1}, \dots, dx^{i'_r}, \frac{\partial}{\partial x^{j'_1}} \cdots \frac{\partial}{\partial x^{j'_s}}\right).$$

Since $x^{i'}$ can be treated as functions of x^i, substituting in the above for $\partial/\partial x^{i'}$ and $dx^{i'}$ gives for the transformed components of the tensor T,

$$T^{i'_1 \dots i'_r}{}_{j'_1 \dots j'_s} = T^{i_1 \dots i_r}{}_{j_1 \dots j_s} \frac{\partial x^{i'_1}}{\partial x^{i_1}} \cdots \frac{\partial x^{i'_r}}{\partial x^{i_r}} \frac{\partial x^{j_1}}{\partial x^{j'_1}} \cdots \frac{\partial x^{j_s}}{\partial x^{j'_s}}. \tag{2.26}$$

Thus, for the transformation of the components of a vector V and a one-form ω we get

$$V^{i'} = \frac{\partial x^{i'}}{\partial x^i} V^i, \quad \omega_{i'} = \frac{\partial x^i}{\partial x^{i'}} \omega_i. \tag{2.27}$$

Even when one chooses a general set of basis vectors rather than a choice of a coordinate basis, the formula for transformation of the components of a tensor can be written in a similar manner.

If T is a tensor of type (r, s), the *contraction* of T over a contravariant index and a covariant index is defined to be a tensor $C(T)$ of type $(r-1, s-1)$. For example, if we contract over the first contravariant and covariant indices, this gives

$$C^1_1(T) = T^{ij \dots l}{}_{im \dots n} e_j \otimes \dots \otimes e_l \otimes e^m \otimes \dots \otimes e^n. \tag{2.28}$$

Using the relationships given above for the transformation of components of a tensor under the change of basis vectors, it is again possible to see that the contraction C^1_1 is independent of the basis used, that is, it is invariant under change of coordinates. Similarly, one could contract T over any pair of a contravariant and a covariant indices.

In the space of all tensors of type (r, s) at p, the *addition* of two tensors T and T' is defined as

$$\begin{aligned}(T + T')(\omega^1, \dots, \omega^r, X_1, \dots, X_s) &= T(\omega^1, \dots, \omega^r, X_1, \dots, X_s) \\ &\quad + T'(\omega^1, \dots, \omega^r, X_1, \dots, X_s),\end{aligned}$$

and the *multiplication by a real number* α is defined by

$$(\alpha T)(\omega^1, \dots, \omega^r, X_1, \dots, X_s) = \alpha T(\omega^1, \dots, \omega^r, X_1, \dots, X_s).$$

The operation of *outer product* on two tensors T and T' of type (r,s) and (r',s') could now be defined in terms of their components to give a new tensor $T\otimes T'$ given by,

$$(T\otimes T')^{i_1\ldots i_{r+r'}}{}_{j_1\ldots j_{s+s'}} = T^{i_1\ldots i_r}{}_{j_1\ldots j_s} T'_{j_{s+1}\ldots j_{s+s'}}{}^{i_{r+1}\ldots i_{r+r'}}.$$

This offers a way of constructing new tensors out of the vectors and dual vectors.

A *tensor field* of (r,s) type on M is an assignment of a tensor of the same type at all p in M. Such a tensor field is called C^k *differentiable* if all the components of T are having the same differentiability as functions of coordinates.

Finally, we discuss the symmetry properties of tensors. Suppose T is a $(0,2)$ type tensor. Then it acts on pairs of vectors V,W to produce a real number $T(V,W)=T_{ij}V^iW^j$. Then T is called *symmetric* if the result is the same even when we change the slots for V and W, that is,

$$T(V,W)=T(W,V).$$

If $\{e_i\}$ is a basis for the tangent space, this amounts to the condition $T(e_i,e_j)=T(e_j,e_i)$, which is same as saying that

$$T_{ij}=T_{ji}.$$

Similarly, T is called *antisymmetric* if

$$T_{ij}=-T_{ji}.$$

It is convenient to formulate this in terms of symmetric and antisymmetric parts of T. For a tensor with components T_{ij}, its *symmetric part* is written as,

$$T_{(ij)}=\frac{1}{2!}(T_{ij}+T_{ji}),$$

and its *antisymmetric part* is written as

$$T_{[ij]}=\frac{1}{2!}(T_{ij}-T_{ji}).$$

Then T is called *symmetric* if $T_{(ij)}=T_{ij}$ and it is called *antisymmetric* if $T_{[ij]}=T_{ij}$. In general, for a tensor $T_{i_1,\ldots,i_r}$ of type $(0,r)$, $T_{(i_1\ldots i_r)}$ is defined as the sum over all permutations of indices $i_1,\ldots,i_r$ divided by $r!$ Similarly, $T_{[i_1\ldots i_r]}$ is defined as the alternating sum over all permutations of the indices $i_1,\ldots,i_r$ divided by $r!$ Thus, for example,

$$T^i{}_{[jkl]}=\frac{1}{3!}[T^i{}_{jkl}+T^i{}_{klj}+T^i{}_{ljk}-T^i{}_{kjl}-T^i{}_{lkj}-T^i{}_{jlk}].$$

In general, a tensor of type (r, s) is called symmetric over a collection of indices if it equals its symmetric part over these indices, and antisymmetric tensors are defined in a similar manner.

2.4 The metric tensor and connection

One desires to have a notion of distance between any two infinitesimally separated points of a space-time manifold. Such distances should locally reduce to those defined by the special theory of relativity that is, those given by a flat metric with an indefinite signature on the Minkowski space-time. This is because special relativity is the theory which has been shown to be valid by experiments and hence must hold at least when confined to local regions in the space-time which correspond to the measurements of space and time intervals at the laboratory scale for an observer. Thus, the distances between events in a space-time need not necessarily be positive.

This is achieved by assuming the existence of an indefinite *metric tensor field* defined globally on M as a $(0, 2)$ type, symmetric, tensor field. Thus, the metric tensor must act on pairs of vectors to produce a number and it is symmetric in its indices. Choosing a coordinate basis this can be written as

$$g \equiv g_{ij} dx^i \otimes dx^j, \tag{2.29}$$

where $g_{ij} = g(\partial/\partial x^i, \partial/\partial x^j)$. If V and W are any two vectors, this gives $g(V, W) = g_{ij} V^i W^j$. This is often written conventionally in the form of an expression giving the distance between two infinitesimally separated points in the space-time as

$$ds^2 = g_{ij} dx^i dx^j. \tag{2.30}$$

For a single vector V, $g(V, V)$ gives the magnitude of V, which is $g_{ij} V^i V^j$.

Another property assumed for the metric tensor is that it is *non-degenerate*, that is, there is no non-zero vector $V \neq 0$ such that $g(V, W) = 0$ for all vectors $W \in T_p$. This amounts to saying that the matrix $[g_{ij}]$ is non-singular and hence there must be an inverse matrix g^{ij} such that

$$g^{ij} g_{jk} = \delta^i{}_k.$$

Hence, the tensors g^{ij} and g_{ij} provide an isomorphism or a unique correspondence between the space of covariant and contravariant vectors in the following sense:

$$X_i = g_{ij} X^j, \qquad X^i = g^{ij} X_j.$$

Similarly, we can also write for a second rank tensor T,

$$T^i{}_j = g^{ik} T_{kj}, \qquad T^j{}_i = g^{jk} T_{ki}, \qquad T^{ij} = g^{ik} g^{jl} T_{kl}.$$

In particular, we have

$$g^{ik}g_{km} = g^i{}_m = \delta^i{}_m,$$

and the Kronecker delta $\delta^i{}_m$ transform as components of a tensor. Thus, $\delta^i{}_m$ and $g^i{}_m$ are identical tensors.

The tensors $T^i{}_j$, $T^j{}_i$, or T^{ij} are to be regarded as representations of the same geometric object because these are uniquely associated tensors. Such an isomorphism between the covariant and contravariant arguments is essentially equivalent to the procedure of 'raising' and 'lowering' of indices as pointed out above. In fact, the multilinear map

$$g : T_p \times T_p \to \mathbb{R}$$

can also be viewed as a linear correspondence from T_p to T_p^* in the sense of the mapping $V \to g(., V)$. The non-degeneracy of the metric tensor implies that this map is one-one and onto and thus g establishes a one-one correspondence between vectors and dual vectors. The components $V_i = g_{ij}V^j$ are the one-form components uniquely associated with the vector components V^j.

Suppose M is an n-dimensional manifold with g being the metric tensor defined on it. Then, at any $p \in M$ one could always choose an orthonormal basis $\{e_i\}$ such that the metric components g_{ij}s have the diagonal form

$$g_{ij} = diag(+1, \ldots, +1, -1, \ldots, -1).$$

If the metric has the form $g_{ij} = (+1, \ldots, +1)$ then it is called *positive definite*. In that case, $g(X, X) = 0$ implies $X = 0$. On the other hand, it is called a *Lorentzian metric* if the form is

$$g_{ij} = diag(+1, \ldots, +1, -1), \tag{2.31}$$

where there are $(n - 1)$ terms with $+1$. This is an indefinite metric in the sense that the magnitude of a non-zero vector could be either positive, negative or zero. Then, $X \in T_p$ is called *timelike*, *null*, or *spacelike*, depending on

$$g(X, X) < 0, \qquad g(X, X) = 0, \qquad g(X, X) > 0. \tag{2.32}$$

An indefinite metric divides the vectors in T_p into three disjoint classes, namely the timelike, null, and spacelike vectors. The null vectors form a cone in the tangent space T_p which separates the timelike vectors from the spacelike vectors (Fig. 7).

When the manifold has dimension four, and when it is equipped with a globally defined Lorentzian metric tensor field, it is called a *space-time manifold*. The *signature* of the metric tensor is defined as the number of

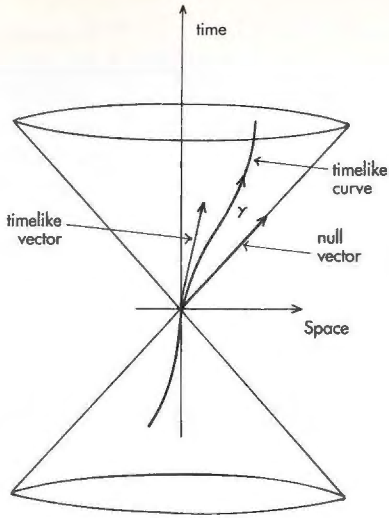

Fig. 7 The null cone at a point p in the space-time manifold. The tangent to curve γ is timelike at all points, which is thus a timelike curve.

its positive eigenvalues minus the number of negative eigenvalues. Thus, a space-time is a four-dimensional differentiable manifold with a Lorentzian metric globally defined which has the signature $+2$.

In fact, in the special theory of relativity, the space-time admits a global coordinate frame covering the entire manifold so that the metric has the form given by (1.2) globally, and the metric coefficients are constants throughout the manifold, which is called the Minkowski space-time. The tangent vector for a particle travelling with a constant velocity less than that of light through a point p in such a space-time is represented by a timelike vector at p. The particle must travel within the future light cone at p which satisfies the equation $g(X, X) = 0$. This equation gives the set of all null vectors at p representing the photon paths. Now, according to the special theory of relativity, no material particles and signals could travel at a velocity more than that of light. Thus, the metric determines the causal structure of space-time in the sense that an event p is causally related to another event q if and only if there is a timelike or null signal between p and q. All such events lie on or within the double cone at p which is defined by the metric tensor in the above manner.

For a non-flat space-time continuum of the general theory of relativity, the metric coefficients are functions of the space-time coordinates and one has to solve for the metric as a solution of the Einstein field equations. As far as the existence of a Lorentz metric on a space-time is concerned, any

C^r paracompact manifold will admit a C^{r-1} Lorentz metric if and only if it admits a nonvanishing C^{r-1} line element field, which is an assignment of a pair of equal and opposite vectors $(V, -V)$ globally on M at each point (see for example, Hawking and Ellis, 1973). Such a line element field is always defined for a non-compact manifold and hence a Lorentz metric always exists for the same. For the reasons explained in Section 2.2, we always take the space-time to be non-compact and without boundary.

Let (M, g) be a space-time and γ be a continuous C^1 curve in M. Then γ is called a *timelike, null,* or *spacelike curve* respectively if the tangent vector to γ is timelike, null, or spacelike respectively at all points of γ. A curve which is either timelike or null is also sometimes called a *non-spacelike curve.* The tangent space magnitudes defined by g, namely,

$$X \to |\, g(X, X)\, |^{1/2},$$

can be related to the magnitudes or distances on the manifold as below. Suppose X is the tangent vector along γ such that $g(X, X)$ has the same sign at all points of $\gamma(t)$. Then the *arc length* between $p = \gamma(t_1)$ and $q = \gamma(t_2)$ along the curve is given by

$$L(\gamma) = s = \int_a^b (|\, g(X, X)\, |)^{1/2} dt. \tag{2.33}$$

The above as well as the relation (2.29) are equivalent to the expression $ds^2 = g_{ij} dx^i dx^j$, which represents the infinitesimal arc length along γ.

In Euclidian spaces one has the notion of parallel transport of any given vector X defined by the condition that in going from a point p to another point q, both the magnitude and the direction of X must not change. If along some curve the magnitude and direction of the tangent vector remain unchanged, such a curve is called a straight line along which the tangent vector is parallel transported. In an Euclidian space, if a vector is parallel transported from a point p to another point q along two different curves, the result will be the same, independent of the path taken. However, this will not be case for a general affine manifold (Fig. 8). For a general differentiable manifold, such a notion of parallel transport of vectors is defined by introducing the concept of a connection on M.

Let X be a vector field on M. First we introduce the notion of a *derivative operator* ∇_X on M which gives the rate of change of vectors or tensor fields along the given vector field X at p for all points of M. If Y is another vector field at p then the operator ∇_X maps Y into a new vector field $Y \to \nabla_X Y$ such that the following conditions are satisfied:

(1) $\nabla_X(\alpha Y + \beta Z) = \alpha \nabla_X Y + \beta \nabla_X Z; \qquad \alpha, \beta \in \mathbb{R}$
(2) $\nabla_{fX+gY} Z = f \nabla_X Z + g \nabla_Y Z;$ for real functions f and g

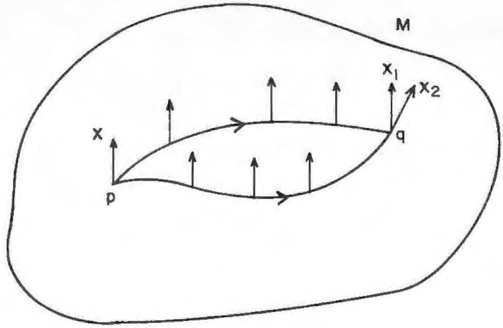

Fig. 8 In a differentiable manifold, the result of parallel transport of a vector along a curve from point p to q in general depends on the path taken.

(3) $\nabla_X(fY) = f\nabla_X Y + YX(f)$

A *connection* ∇ at a point $p \in M$ is a rule which assigns to each vector field X at p a differential operator ∇_X which maps an arbitrary C^r vector field Y at p into a vector field $\nabla_X Y$ such that the conditions (1), (2), and (3) are satisfied (for a general discussion on connections on manifolds, see for example, Hicks (1965)). Thus, ∇Y, called the *covariant derivative of* Y is defined as a type $(1,1)$ tensor field which gives a vector $\nabla_X Y$ when contracted with the vector X. In such a case, the condition (3) above implies

$$\nabla(fY) = df \otimes Y + f\nabla Y. \tag{2.34}$$

A C^r *connection* ∇ on a C^k manifold $(k \geq r+2)$ is a rule assigning a connection ∇ to each $p \in M$ such that if Y is a C^{r+1} vector field, then ∇Y is a C^r tensor field of type (1,1). We can write

$$\nabla Y = Y^i{}_{;j} e^j \otimes e_i. \tag{2.35}$$

Here $Y^i{}_{;j}$ is often called the *covariant derivative* of the vector Y^i. This is completely defined by the n^3 *connection coefficients* $\Gamma^i{}_{jk}$ which are defined in the following manner by choosing the vector fields X and Y to be the basis vector fields:

$$\nabla_{e_j} e_k \equiv \Gamma^i{}_{jk} e_i. \tag{2.36}$$

It is not difficult to see that the above is equivalent to the condition

$$\langle e^i, \nabla_{e_j} e_k \rangle = \Gamma^i{}_{jk}. \tag{2.37}$$

Thus, in a coordinate basis we have

$$\langle dx^i, \nabla_{\partial/\partial x^j} \left(\frac{\partial}{\partial x^k} \right) \rangle = \Gamma^i{}_{jk}.$$

Consider now the vector $\nabla_X Y$. Defining

$$\nabla_{\partial/\partial x^i} Y \equiv \nabla_i Y,$$

using the rules defining the connection given above, and the relation

$$X(f) = X^i \frac{\partial}{\partial x^i}(f) = X^i \frac{\partial f}{\partial x^i}, \tag{2.38}$$

we obtain

$$\nabla_X Y = X^i \left(\frac{\partial Y^k}{\partial x^i} + \Gamma^k{}_{ij} Y^j \right) \left(\frac{\partial}{\partial x^k} \right). \tag{2.39}$$

Comparing this with eqn (2.35), we can write

$$\nabla_X Y = Y^k{}_{;i} X^i \left(\frac{\partial}{\partial x^k} \right)$$

where

$$Y^k{}_{;i} \equiv \frac{\partial Y^k}{\partial x^i} + \Gamma^k{}_{ij} Y^j. \tag{2.40}$$

It can be seen that the components of the vector $\nabla_X Y$ are given as $Y^k{}_{;i} X^i$.

Let us define

$$Y^i{}_{,j} \equiv \frac{\partial Y^i}{\partial x^j}.$$

Then, taking the transformation of coordinates $\{x^i\} \to \{x^{i'}\}$ when the basis vectors transform as $e_i \to e_{i'}$, it can be seen that $Y^i{}_{,j}$ does not transform like the components of a tensor. Similarly, consider the connection coefficients in the new coordinate system, which are given by

$$\Gamma^{k'}{}_{i'j'} = \langle e^{k'}, \nabla_{e_{i'}} e_{j'} \rangle.$$

Transforming the dashed vectors to the original coordinate system and using the condition (2) and (3) above gives in a coordinate basis,

$$\Gamma^{k'}{}_{i'j'} = \frac{\partial x^{k'}}{\partial x^k} \left(\frac{\partial x^i}{\partial x^{i'}} \frac{\partial x^j}{\partial x^{j'}} \Gamma^k{}_{ij} + \frac{\partial^2 x^k}{\partial x^{i'} \partial x^{j'}} \right). \tag{2.41}$$

It follows that because of the presence of the second derivative terms in the above, the coefficients $\Gamma^i{}_{jk}$ also do not transform like the components of a tensor. Consider, however,

$$\nabla_X Y = (Y^i{}_{;j} X^j) \left(\frac{\partial}{\partial x^i} \right) = (Y^{i'}{}_{;j'} X^{j'}) \left(\frac{\partial}{\partial x^{i'}} \right)$$

which implies

$$Y^i{}_{;j}X^j = Y^{i'}{}_{;j'}\frac{\partial x^{j'}}{\partial x^j}\frac{\partial x^i}{\partial x^{i'}}X^j. \tag{2.42}$$

Since the above is true for any arbitrary vector X^j, it follows that $Y^i{}_{;j}$ are components of a tensor.

Further, if $\Gamma^i{}_{jk}$ and $\bar{\Gamma}^i{}_{jk}$ are components of two different connections on M, then it is not difficult to see, using the coordinate transformations, that the quantities

$$C^i{}_{jk} = \bar{\Gamma}^i{}_{jk} - \Gamma^i{}_{jk}$$

are components of a tensor.

Given a connection ∇ on M, the *torsion tensor* T is defined by the relation

$$T(X,Y) = \nabla_X Y - \nabla_Y X - [X,Y]. \tag{2.43}$$

Writing the components,

$$T(X,Y) = (\Gamma^i{}_{jk} - \Gamma^i{}_{kj})X^j Y^k e_i.$$

This is a type $(1,2)$ tensor which has the components

$$T^i{}_{jk} = \Gamma^i{}_{jk} - \Gamma^i{}_{kj}.$$

A connection is called *symmetric* when the torsion tensor vanishes, that is,

$$\Gamma^i{}_{jk} = \Gamma^i{}_{kj},$$

or $[X,Y] = \nabla_X Y - \nabla_Y X$. We shall always work with symmetric connections and assume the torsion tensor to be vanishing.

The notion of connection can be generalized to arbitrary tensor fields to obtain a tensor $\nabla_X T$ of type (r,s) for any given tensor T of (r,s) type by assuming first that ∇ is linear and obeys the Leibnitz rule. That is,

$$\nabla_X(\alpha S + \beta T) = \alpha\nabla_X S + \beta\nabla_X T; \qquad \alpha,\beta\in\mathbb{R}$$

and

$$\nabla_X(S\otimes T) = \nabla_X S\otimes T + S\otimes\nabla_X T$$

for any vector field X. Further, ∇ must agree with our usual notion of directional derivative, that is,

$$\nabla_X f = \langle df, X\rangle = Xf = X^i\frac{\partial f}{\partial x^i}. \tag{2.44}$$

Finally, ∇ must commute with contractions, that is,

$$(\nabla_a T)^{i_1 \dots l \dots i_r}{}_{j_1 \dots l \dots j_s} = \nabla_a T^{i_1 \dots l \dots i_r}{}_{j_1 \dots l \dots j_s}. \tag{2.45}$$

As earlier, we can write

$$\nabla_X T = T^{i_1 \dots i_r}{}_{j_1 \dots j_s;a} X^a e_{i_1} \otimes \dots \otimes e_{i_r} \otimes e^{j_1} \otimes \dots \otimes e^{j_s}, \tag{2.46}$$

with

$$\nabla_X T^{i_1 \dots i_r}{}_{j_1 \dots j_s} = T^{i_1 \dots i_r}{}_{j_1 \dots j_s;a} X^a.$$

Now, by considering the expansion for $\nabla_i(e_j \otimes e^k)$ it is not difficult to see that

$$\nabla_a e^i = -\Gamma^i{}_{ac} e^c, \tag{2.47}$$

and if ω is a one-form then,

$$\nabla_{e_j} \omega = \omega_{k;j} e^k,$$

with

$$\omega_{k;j} \equiv \frac{\partial \omega_k}{\partial x^j} - \Gamma^i{}_{jk} \omega_i. \tag{2.48}$$

In general, we can write for the covariant derivative of a tensor T,

$$\begin{aligned} T^{i_1 \dots i_r}{}_{j_1 \dots j_s;a} = \frac{\partial T^{i_1 \dots i_r}{}_{j_1 \dots j_s}}{\partial x^a} &+ \sum_m \Gamma^{i_m}{}_{he} T^{i_1 \dots e \dots i_r}{}_{j_1 \dots j_s} \\ &- \sum_n \Gamma^e{}_{h j_n} T^{i_1 \dots i_r}{}_{j_1 \dots e \dots j_s}. \end{aligned} \tag{2.49}$$

Finally, we note that given a Lorentzian metric tensor on M, the condition $\nabla_X g = 0$ defines a unique torsion-free connection on M. Then,

$$(\nabla_X g)_{ij} = g_{ij;k} X^k = 0, \tag{2.50}$$

which implies that

$$g_{ij;k} = 0.$$

In such a case, the parallel transport of vectors must preserve the scalar product defined by the metric tensor g and the connection coefficients $\Gamma^i{}_j k$ are determined in terms of the first derivatives of the metric components. Since all the information on space-time structure is supposed to be contained in the ten metric functions g_{ij}, this is reasonable to expect. One way to see this is the following. Using eqn (2.49), we can write for the covariant derivative of the metric,

$$g_{ij;k} = \frac{\partial g_{ij}}{\partial x^k} - \Gamma^m{}_{il} g_{mj} - \Gamma^m{}_{jl} g_{mi}.$$

Now, using the condition $g_{ij;k} = 0$ and defining

$$g_{mj} \Gamma^m{}_{il} = \Gamma_{jil},$$

the above can be written as

$$g_{ij,k} \equiv \frac{\partial g_{ij}}{\partial x^k} = \Gamma_{jil} + \Gamma_{ijl}.$$

Using the above and the symmetry property of the connection, we get

$$\Gamma_{jil} = \tfrac{1}{2}(g_{ji,m} - g_{mj,i} + g_{im,j}).$$

This result can be seen by specializing to the frame of free fall as well. In such a frame, all the connection coefficients vanish and the metric is locally that of the special theory of relativity. Then, $g_{ij} = \eta_{ij}$ and the partial derivatives of g_{ij} vanish. Thus, from the above equation for $g_{ij;k}$, we again recover $g_{ij;k} = 0$. This, being a tensor equation, must hold in all frames in general and we can again proceed as earlier.

2.5 Non-spacelike geodesics

In Euclidian spaces, given any two points, the line of shortest distance between them is the straight line joining the two points along which the tangent vector does not change either the direction or magnitude, that is, the tangent vector is parallel transported. Now, let $\gamma(t) : \mathbb{R} \to M$ be a C^1 curve in M. If T is a $C^r (r \geq 0)$ tensor field on M, then the *covariant derivative of T along $\gamma(t)$* is defined as

$$\frac{DT}{\partial t} = T^{i\ldots l}{}_{k\ldots m;h} X^h,$$

where X is the tangent vector to $\gamma(t)$. Then γ is called a *geodesic* if the tangent vector to γ is parallel transported along it. In other words, if X denotes the tangent vector field along γ, then it is required that $\nabla_X X$ is proportional to X, that is, there exists a function f such that

$$\nabla_X X = f X. \tag{2.51}$$

Writing the components, this implies that $(X^i{}_{;j}X^j)e_i = fX^ie_i$ always holds, and so one has $X^i{}_{;j}X^j = fX^i$ along the geodesic curve. However, it is always possible to reduce the function f to zero by a suitable choice of the curve parameter t along the geodesic γ, and the equation for the geodesic is written as

$$X^i{}_{;j}X^j = 0, \tag{2.52}$$

where X^i are the components of the tangent vector to the geodesic. Such a parameter t is called an *affine parameter* along γ, which is called an *affinely parametrized geodesic.* If $\{x^i\}$ denotes a local coordinate system, the components X^i are written as $X^i = dx^i/dt$ and the above equation for geodesics is written as

$$\frac{d^2x^i}{dt^2} + \Gamma^i{}_{jk}\frac{dx^j}{dt}\frac{dx^k}{dt} = 0. \tag{2.53}$$

The affine parameter along the geodesic is determined up to an additive and multiplicative constant. Thus, if t is an affine parameter, then so is $t' = at + b$ and we again have $X^i{}_{;j}X^j = 0$. Here $b \neq 0$ gives a new choice of the initial point $\gamma(0)$ and $a \neq 0$ implies a renormalization of the vector X.

As in the case of general curves, such a geodesic in a space-time (M, g) is called either timelike, spacelike, or null, depending on whether the tangent vector to the geodesic is timelike, spacelike, or null respectively. We shall be interested here in the timelike or null geodesics as they represent the paths of particles or photons in the space-time. Since the tangent vector to the geodesic is parallel transported, a timelike or null geodesic remains timelike or null always and could not become spacelike. In a Riemannian manifold with a positive definite metric, such geodesics give the curves of shortest distance between its points. However, in a space-time with a Lorentzian metric, the non-spacelike geodesics maximize the distance between the points as defined by eqn (2.33). If there is a timelike geodesic between the points p and q, there is no shortest distance geodesic between these points because, by introducing null geodesic pieces, one could join these points by curves of arbitrary small lengths. On the other hand, any maximal length curve between p and q must necessarily be a timelike geodesic.

The geodesic equations (2.53) are n equations in n-variables x^i with $i = 1, \ldots, n$. Thus the existence theorems for the differential equations ensure that given x^i and dx^i/dt, that is, given any initial point p and the value of the tangent vector X^i, there exists a unique geodesic through p with this value of the tangent. This can be used to define the *exponential map* $E_p : T_p \to M$ from the tangent space at p into the space-time as below. Under this map any given tangent vector X^i in T_p is mapped to a point in M a unit affine parameter distance away along the unique geodesic determined by p and X^i. It is clear that the exponential map may not be defined on all of T_p because all the geodesics in M passing through p may not extend to all the values of the affine parameter. In such a case M is called *geodesically incomplete.* On the other hand, if the exponential map is defined on all of T_p for all points p, then M is called *geodesically complete*, in which case all geodesics in M extend for all values of their affine parameter. Also, the map E_p may not be one-one because the geodesics might cross each other. However, it can be shown (Bishop and Critendon, 1964) that for a sufficiently small neighbourhood N_p of p there is a neighbourhood of the origin in T_p which is diffeomorphically mapped onto N_p by the exponential map which is one-one and well defined on this neighbourhood. In this case the exponential map can be used to define the *Riemannian normal coordinates* on the neighbourhood N_p of p. Since T_p is an n-dimensional vector space equivalent to $\mathbb{R}^n$, the coordinates of any $r \in N_p$ can be chosen to be the n-coordinates of the vector X_p which is mapped onto it. This coordinate system has the property that the geodesics are mapped into straight lines and the connection coefficients vanish at p. Thus, such a coordinate system turns out to be quite convenient for calculations at the point p. The neighbourhood N_p could have a further property that any two points in it can be joined by a unique geodesic contained totally within N_p. Such a neighbourhood of p is called a *convex normal neighbourhood.*

The geodesic equations (2.52) were derived from the requirement that the tangent vector to the curve must be parallel transported. Now, the length of a curve, as defined by the metric tensor, is given by eqn (2.33). If we impose the condition that the curve must extremize the length, namely that, $\delta l = 0$, and work out

$$\delta l = \delta \int_a^b \mid g_{ab} \frac{dx^a}{dt} \frac{dx^b}{dt} \mid^{1/2} dt = 0,$$

using the variational methods, it turns out that the resulting equations are precisely the geodesic equations (2.53). It follows that the geodesics extremise the lengths of curves between any two space-time points. In particular, if the points p and q are timelike related and if there is a maximum length timelike curve from p to q, that curve must be a timelike geodesic. (We shall show in Chapter 4 that there are situations in which there would be no maximal length timelike curve between two timelike related points p and q.) Thus, in most situations, the useful way to work out geodesics equations of the space-time is to choose the Lagrangian as

$$L = \tfrac{1}{2} g_{ab} \dot{x}^a \dot{x}^b$$

and to write down the Lagrange equations which are the equations of the space-time geodesics. Then, by comparison with eqn (2.53), it is possible to evaluate the quantities $\Gamma^i{}_{jk}$ for the space-time as well.

Consider now a *conformal transformation* of the metric g_{ij}, namely that given by

$$g_{ij} \to \bar{g}_{ij} = \Omega^2 g_{ij}, \tag{2.54}$$

where Ω is a real-valued function of the coordinates x^i such that $0 < \Omega < \infty$ and it satisfies a suitable differentiability condition. Such a transformation of the space-time metric preserves the angles between the vectors and the ratios of their magnitudes (but not their lengths). Clearly, any null vector with respect to g remains null with respect to $\bar{g}$ as well and so it can be shown that all the null geodesics with respect to g remain null curves in general (which need not necessarily be geodesics) with respect to $\bar{g}$. Thus, the null cone at each point, and as a result the causal structure of the space-time, is preserved under a conformal transformation of the metric tensor. Of course, the timelike geodesics of the two space-times (M, g) and $(M, \bar{g})$ will not be the same, however, if an event p is chronologically related to another event q in (M, g), the same relation will hold in $(M, \bar{g})$ as well as a result of the invariance of the null cones under conformal transformations. In general, if there is a timelike geodesic from a point p to q in (M, g), there will be a timelike curve between these points in $(M, \bar{g})$.

On the other hand, if $\bar{g}$ is any other metric which preserves the causal structure of the original space-time (M, g), then $\bar{g}$ must be conformal to g. To see this, write the metric g as $g_{ij} = diag(-1, +1, +1, +1)$ by a suitable choice of a basis $\{e_i\}$, in which case the basis vectors will be orthogonal to each other, each having a unit magnitude. The equation

$$g(e_0 + \lambda e_1, e_0 + \lambda e_1) = g(e_0, e_0) + 2\lambda g(e_0, e_1) + \lambda^2 g(e_1, e_1) = 0,$$

will have two real roots $\lambda = \pm 1$, and hence the vectors $e_0 \pm e_\alpha, \quad \alpha = 1, 2, 3$ are null vectors with respect to g. Since $\bar{g}$ preserves null cones, these vectors are null with respect to $\bar{g}$ as well which implies that e_0 is orthogonal to e_α and these vectors have the same magnitude with respect to $\bar{g}$ also. Next, the equation,

$$g(e_0 + \lambda(e_\alpha + e_\beta), e_0 + \lambda(e_\alpha + e_\beta)) = 0,$$

for $\alpha \neq \beta$ has a root $\lambda = 1/\sqrt{2}$. Hence, $e_0 + 1/\sqrt{2}(e_\alpha + e_\beta)$ is null with respect to g which implies that the vectors $e_\alpha, \quad \alpha = 1, 2, 3$ are mutually orthogonal with respect to $\bar{g}$ also. Thus, the orthogonal basis $\{e_i\}$ for g is mapped into an orthogonal basis for $\bar{g}$ where all basis vectors have the same magnitude and hence the components $\bar{g}_{ij}$ are the same as g_{ij} at p apart from a constant multiple. Thus, the metrics g and $\bar{g}$ must be conformally related.

Such conformal transformations of the metric tensor alter the geodesic completeness properties of the space-time, which we shall discuss in Chapters 4 and 9. However, in order to compare the non-spacelike geodesics in two conformally related space-times (M, g) and $(M, \bar{g})$, let ∇ and $\bar{\nabla}$ denote the connections on these space-times respectively. Let X^a be the tangent to an affinely parametrized non-spacelike geodesic γ with respect to ∇ which satisfies $\nabla_X X = 0$. Then we have

$$\bar{\nabla}_X X - \nabla_X X = [X^i(\bar{\Gamma^j}_{ik} - \Gamma^j{}_{ik})X^k]e_j,$$

where e_j are basis vectors. However, as noted in the previous section, in the case of a metric connection the coefficients $\Gamma^i{}_{jk}$ are uniquely determined in terms of the metric components. Using this, and the relation $\bar{g}_{ij} = \Omega^2 g_{ij}$, we can write

$$(\bar{\nabla}_X X)^j = X^i C^j{}_{ik} X^k = 2X^j X^k \nabla_k(\ln\Omega) - (g_{ik}X^i X^k) g^{jl} \nabla_l(\ln\Omega).$$

It is clear that γ is no longer a geodesic with respect to $(M, \bar{g})$. However, in the case of null geodesics, the above is just the geodesic eqn (2.51) with the function $f = X^k \nabla_k(\ln\Omega)$, where the curve parameter is no longer affinely parametrized. Thus, one could state that the null geodesics of two conformally related space-times are the same apart from the affine parameters which are related by

$$\frac{d\bar{v}}{dv} = \text{const.}\,\Omega^2.$$

It is useful to introduce at this stage the notion of a hypersurface in a space-time. In the Minkowski space-time, the surface $t = 0$ is a three-dimensional surface with the time direction always normal to it. Any other surface $t =$ const. is also a spacelike surface in this sense. In general, let S be an $(n-1)$-dimensional manifold. If there exists a C^∞ map $\phi : S \to M$ which is locally one-one (that is, there is a neighbourhood N for every $p \in S$ such that ϕ restricted to N is one-one) and ϕ^{-1} is also C^∞ as defined on $\phi(N)$, then $\phi(S)$ is called an *immersed submanifold* of M. If ϕ is globally one-one, then $\phi(S)$ is called an *embedded submanifold of M*. One also might require ϕ to be a homeomorphism with the induced topology on $\phi(S)$ from M. Lower dimensional embedded submanifolds in M represent well-behaved surfaces in the space-time. A *hypersurface* S of any n-dimensional manifold M is defined as an $(n-1)$-dimensional embedded submanifold of M. We denote by V_p the $(n-1)$-dimensional subspace of T_p of the vectors tangent to S at any $p \in S$. It follows that there exists a vector $n^a \in T_p$, which is unique up to the scale and is orthogonal to all the vectors in V_p. This is called the *normal* to S at p. If the magnitude of n^a is either positive or negative at all points of S without changing the sign, then n^a could be normalized so that $g_{ab}n^a n^b = \pm 1$. If $g_{ab}n^a n^b = -1$, then the normal vector is timelike everywhere and S is called a *spacelike hypersurface.* If the normal is spacelike everywhere on S with a positive magnitude, S is called a *timelike hypersurface.* Finally, S is a *null hypersurface* if the normal n^a is null at S.

We finally point out here how the timelike geodesics in the space-time could be used to define the *synchronous coordinate system* in the neighbourhood of a spacelike hypersurface in the space-time, which we describe below. Let ∇_a be the metric connection in the space-time which satisfies $\nabla_a g_{bc} = g_{bc;a} = 0$. Suppose now S is a spacelike hypersurface in a space-time M. For every $p \in S$, let γ be the unique timelike geodesic with tangent n^a, that is, the congruence of these curves at points of S is orthogonal to S. Then, in the neighbourhood of that portion of S we assign the coordinates $q \to (x^1, \ldots, x^{n-1}, t)$ for any point q in future of p along γ; where t is the parameter along γ and $x^1, \ldots, x^{n-1}$ are the spatial coordinates of p. In particular, if the geodesics in the congruence are parametrized by the proper time t with the magnitude of tangent given by -1 along γ, the spacelike surfaces are given as $\{t = \text{const.}\}$ surfaces. We can label S as the $\{t = 0\}$ spacelike surface.

The synchronous coordinates so constructed have the important property that when the congruence $\{\gamma\}$ is orthogonal to S_0, it will also be orthogonal to subsequent surfaces S_t given by $t = \text{const}$. Clearly, $\{\gamma\}$ is orthogonal to S_0 by construction. To see that this is true for any S_t for t within the domain of construction, let X^a be any basis vector for the tangent space at a point of S_t. Then,

$$n^b \nabla_b (n_a X^a) = (n^b \nabla_b n_a) X^a + n^b n_a \nabla_b X^a.$$

Since n^b is tangent vector of the geodesic, the above equals $n_a n^b \nabla_b X^a$. But X and n are coordinate vectors, implying $\nabla_n X = \nabla_X n$, that is,

$$n^a \nabla_a X^b = X^a \nabla_a n^b. \tag{2.55}$$

This implies

$$n^b \nabla_b (n_a X^a) = n_a X^b \nabla_b n^a = \tfrac{1}{2} X^b \nabla_b (n^a n_a) = 0$$

because $n^a n_a = -1$. Thus, $n_a X^a = 0$ in the future of S_0 in the domain of validity of the synchronous coordinate system.

2.6 Diffeomorphisms and Killing vectors

Let ϕ be a C^∞ map from a space-time M to another space-time N, such that ϕ makes a unique assignment of a point in N for each point in M and if $p \in M$ then coordinates of $\phi(p)$ are C^∞ functions of the coordinates of p. Such a map ϕ naturally induces a *pull back map* ϕ^* of all functions on N to functions on M defined as below. If f is any function on N then,

$$(\phi^* f)(p) = f(\phi(p)). \tag{2.56}$$

The map ϕ also induces another map ϕ_* of all the tangent vectors of T_p into the tangent vectors of $T_{\phi(p)}$ defined as below. If V is a tangent vector at p to a curve $\gamma(t)$ passing through p, then $\phi_* V$ is tangent to the image curve $\phi(\gamma(t))$ in N at $\phi(p)$. Thus

$$(\phi_* V)(f) = V(f \circ \phi) = V(\phi^* f), \tag{2.57}$$

where f is any function on N and $V(f)$ denotes the directional derivative. One can also pull back one-forms ω on N into M using ϕ^*, which is defined by the relation

$$(\phi^* \omega)_i V^i = \omega_i (\phi_* V)^i. \tag{2.58}$$

In fact, the maps ϕ^* and ϕ_* can be extended to carry tensors of the (r,s) type between M and N when ϕ is one-one.

Amongst the set of all possible mappings between two space-time manifolds, the diffeomorphisms have a special significance. It can be seen that if $\phi : M \to N$ is a diffeomorphism then M and N have identical manifold structures and the solutions $(M, T^{(i)})$ and $(N, \phi^* T^{(i)})$ have identical physical properties, where $T^{(i)}$s denote tensor fields on M. When the tensor fields are not diffeomorphically related, the space-times have different and physically non-equivalent properties. In this sense, the space-time diffeomorphisms represent the gauge freedom of general relativity theory or for any space-time theory formulated in terms of tensor fields on a space-time manifold. Thus, diffeomorphically related space-times may be considered to be equivalent for all practical purposes.

We shall now confine our attention to the maps ϕ which are diffeomorphisms on M that is, ϕ is a one-one, onto map from M to M such that both ϕ and ϕ^{-1} are C^∞ maps. Then, if $T^{i_1 \ldots i_r}{}_{j_1 \ldots j_s}$ is a tensor of type (r,s) at p, we define the 'carry over' tensor $[\phi_* T]^{i_1 \ldots i_r}{}_{j_1 \ldots j_s}$ at $\phi(p)$ by

$$[\phi_* T]^{i_1 \ldots i_r}{}_{j_1 \ldots j_s} (\omega_1)_{i_1} \ldots (\omega_r)_{i_r} (V^1)^{j_1} \ldots (V^s)^{j_s}$$

$$= T^{i_1 \ldots i_r}{}_{j_1 \ldots j_s} (\phi^* \omega_1)_{i_1} \ldots (\phi^* \omega_r)_{i_r} [(\phi^{-1})_* V^1]^{j_1} \ldots [(\phi^{-1})_* V^s]^{j_s}, \tag{2.59}$$

where $\omega_1, \ldots, \omega_r$ and $V^1, \ldots, V^s$ are one-forms and vectors respectively at $\phi(p)$. The following relations hold for diffeomorphisms:

$$\phi_* = (\phi^{-1})^*, \qquad \phi^* = (\phi^{-1})_*, \tag{2.60}$$

and any one of these maps can carry over tensors in M.

It is possible to view diffeomorphisms locally in terms of local coordinate transformations. Suppose $\phi(p) = q$ and U and V are disjoint coordinate neighbourhoods of p and q respectively such that $\phi^{-1}(V) = U$. Let $\{x^i\}$ and $\{y^i\}$ be local coordinate systems in U and V. We use ϕ to set up a new coordinate system $\{x^{i'}\}$ in the neighbourhood of p, given for any r in U by

$$x^{i'}(r) = y^i(\phi(r)).$$

Thus, the effect of ϕ now is to induce a coordinate transformation $x^i \to x^{i'}$ at p where as the tensor fields at p are left invariant. Here, the components of any tensor field T at p in the $x^{i'}$ system will be the same as the components of $\phi_* T$ at $\phi(p)$ as discussed above. In this sense the local and coordinate-free approaches are equivalent.

For a diffeomorphism $\phi : M \to M$ and for any tensor field T on M, we can compare the tensors T and $\phi_*(T)$. If

$$T = \phi_* T, \tag{2.61}$$

then the map ϕ is called a *symmetry transformation* for the tensor field T. In case of the metric tensor, such a symmetry transformation with the property

$$(\phi_* g)_{ij} = g_{ij}$$

is called an *isometry*. Thus, isometries on M are diffeomorphisms which leave the metric tensor invariant and so the distance measurements in M also will be invariant under ϕ. It is of interest to learn when a quantity does not change under coordinate transformations because this reveals the information about the symmetries of the underlying manifold. For example, the continuous translations and rotations in the Euclidian space leave the distance between points invariant. Viewing the diffeomorphisms as inducing a change of coordinate as indicated above, a coordinate transformation $x^i \to x^{i'}$ is an isometry leaving the metric invariant if we have

$$g_{i'j'}(x) = g_{ij}(x),$$

for all coordinate systems x^i. Thus the functional form of the transformed metric functions remain unchanged. Using this, the coordinate change $x^i \to x^{i'}$ will be an isometry provided

$$g_{ij}(x) = \frac{\partial x^{m'}}{\partial x^i} \frac{\partial x^{n'}}{\partial x^j} g_{m'n'}(x')$$

For a given vector field X^i, one could now consider the special case of an infinitesimal coordinate change given by,

$$x^{i'} = x^i + \epsilon X^i(x)$$

where ϵ is an arbitrary small quantity. Differentiating the above expression and working to the first order in ϵ to work out the partial derivatives in the above expression for $g_{ij}(x)$, one can show that the condition for the coordinate transformation to be an isometry translates to the equations $X_{i;j} + X_{j;i} = 0$. Thus, we conclude that any vector field X^i satisfying this equation generates an infinitesimal isometry for the space-time. Using such isometries, one could build up a finite transformation by integrating over an infinite sequence of such infinitesimal changes.

In general, such symmetries in the space-time can be characterized by the existence of Killing vector fields in M. To discuss this, we first introduce here the notion of the Lie derivative. Let X be a vector field in M. Then, the existence theorems in the theory of differential equations imply that there exists a unique maximal integral curve $\gamma(t)$ through each $p \in M$ such that

$$\gamma(0) = p, \qquad \left(\frac{\partial}{\partial t}\right)_{|t=0} = X_p.$$

Thus, the tangent vector at the point $\gamma(t)$ is $X_{|\gamma(t)}$. Next, for each $q \in M$ there exists an open neighbourhood U containing q and an $\epsilon > 0$ such that X defines a one-parameter family of diffeomorphisms $\phi_t : U \to M$ for $| t | < \epsilon$ obtained by taking each $p \in U$ a parameter distance t along the integral curves of X. These diffeomorphisms form a one-parameter local group given by the rules

$$\phi_{t+s} = \phi_t \circ \phi_s \quad \text{for } | t |, | s |, | t+s | < \epsilon, \quad \phi_{-t} = (\phi_t)^{-1}, \quad \phi_0(p) = p. \tag{2.62}$$

These diffeomorphisms map any tensor T at p as

$$T \to \phi_{t*}T \mid_{\phi_t(p)} .$$

The *Lie derivative* of a tensor field T along X is now defined as

$$L_X T \mid_p = \lim_{t\to 0} \frac{1}{t}[(\phi_{-t})_* T - T]. \tag{2.63}$$

It is clear that the Lie derivative of a tensor T is again a tensor of the same type. Given a derivative operator ∇_a on M it is possible to write down the components of this (see for example, Wald, 1984). If X and Y are any two vector fields then the Lie derivative of Y along the vector X is given by

$$(L_X Y)f = X(Yf) - Y(Xf) = [X, Y]. \tag{2.64}$$

In a coordinate basis, the components of this Lie derivative are then given as

$$(L_X Y)^a = [X, Y]^a = X^b \frac{\partial Y^a}{\partial x^b} - Y^b \frac{\partial X^a}{\partial x^b}. \tag{2.65}$$

For a connection with a vanishing torsion,

$$T(X, Y) = \nabla_X Y - \nabla_Y X - [X, Y] = 0.$$

This gives, for the components of the Lie derivative,

$$(L_X Y)^a = [X, Y]^a = Y^a{}_{;b} X^b - X^a{}_{;b} Y^b = X^b \nabla_b Y^a - Y^b \nabla_b X^a. \tag{2.66}$$

The Lie derivative has the linearity property

$$L_X(aY + bZ) = aL_X Y + bL_X Z,$$

and it also satisfies the Leibnitz property which is the usual product rule for differentiation. That is, if Y is a vector field and T is a second rank tensor, then

$$L_X(Y^a T_{bc}) = Y^a(L_X Z_{bc}) + (L_X Y^a) Z_{bc}.$$

The Lie derivative commutes with contractions in the sense that,

$$g^a{}_b L_X T^b{}_a = L_X T^a{}_a.$$

In particular, for the metric tensor g,

$$(L_X g)_{ab} = \nabla_a X_b + \nabla_b X_a, \tag{2.67}$$

where ∇_a is the connection compatible with the metric, that is,

$$\nabla_a g_{bc} = 0.$$

Alternatively, we can write the above as

$$(L_X g)_{ab} = X_{b;a} + X_{a;b}. \tag{2.68}$$

Suppose now that the local one-parameter group of diffeomorphisms ϕ_t generated by a vector field X is a group of isometries preserving the metric tensor. Then, X is called a *Killing vector field* and we have

$$L_X g = \lim_{t \to 0} \frac{1}{t}[g - (\phi_{-t})_* g] = 0. \tag{2.69}$$

Thus, the necessary and sufficient condition for X to be a Killing vector field is that the Lie derivative of g along X vanishes. Next, since $(L_X g)_{ab} = X_{b;a} + X_{a;b}$ holds for a Killing vector field, we get

$$X_{a;b} + X_{b;a} = 0, \tag{2.70}$$

which is called the *Killing equation.*

The existence of Killing vector fields in a space-time play an important role in generating exact solutions of Einstein's equations. In order to obtain any specific solution, one normally considers a space-time admitting various symmetries which are characterized by the existence of different Killing vectors. An important implication of the existence of a Killing vector X for a space-time is the following. Since the one-parameter group of diffeomorphisms generated by X are all isometries, if a small but finite displacement is made of all space-time points by means of ϕ_t at each point of M then the resulting space-time structure is the same as the original space-time; that is, the space-time is invariant under such a change.

To consider another implication, let X_a be a Killing vector field and γ be a geodesic in M with tangent vector u^a. Then we show that $X_a u^a$ is constant along γ. For this, consider

$$u^b \nabla_b (X_a u^a) = u^b X_a \nabla_b u^a + u^b u^a \nabla_b X_a. \tag{2.71}$$

Now, the Killing equation $\nabla_a X_b + \nabla_b X_a = 0$, when contracted with $u^a u^b$, implies that the second term in eqn (2.71) must be zero. The first term also vanishes because of the geodesic equation, which proves the result. In general relativity, timelike geodesics represent the particle trajectories and the null geodesics represent the light rays. The above result then implies that every one-parameter group of symmetries gives rise to a conserved quantity for particles and light rays. Such conserved quantities could be physically important, and at the same time they also help in integrating geodesic equations.

2.7 Space-time curvature

The measure of the curvature for any given space-time is exhibited in the non-commutation of the tangent vectors when parallel transported along different curves to arrive at the same space-time point. This is given by the *Riemann curvature tensor* which is defined as a type $(1, 3)$ tensor, $R : T_p^* \times T_p \times T_p \times T_p \to \mathbb{R}$. In a coordinate basis one could write the Riemann tensor as

$$R = R^i{}_{jkl} e_i \otimes e^j \otimes e^k \otimes e^l. \tag{2.72}$$

If we define the vector $R(X,Y)Z$ as

$$R(X,Y)Z \equiv \nabla_X(\nabla_Y Z) - \nabla_Y(\nabla_X Z) - \nabla_{[X,Y]}Z, \tag{2.73}$$

then the components of the Riemann tensor are given by

$$R^i{}_{jkl} = \langle e^i, R(e_k, e_l)e_j \rangle. \tag{2.74}$$

Working out the components gives

$$R(X,Y)Z = R^i{}_{jkl}\frac{\partial}{\partial x^i}X^k Y^l Z^j. \tag{2.75}$$

Now, in order to evaluate eqn (2.74), note that

$$[\nabla_X(\nabla_Y Z)]^i = \nabla_X(Z^i{}_{;j}Y^j) = Z^i{}_{;jk}Y^j X^k + Z^i{}_{;j}Y^j{}_{;k}X^k. \tag{2.76}$$

Similarly we have

$$[\nabla_Y(\nabla_X Z)]^i = Z^i{}_{;jk}X^j Y^k + Z^i{}_{;j}X^j{}_{;k}Y^k. \tag{2.77}$$

Finally, we have

$$\begin{aligned} -\nabla_{[X,Y]}Z &= -\nabla_{(Y^i{}_{;j}X^j - X^i{}_{;j}Y^j)(\partial/\partial x^i)}Z \\ &= -Z^k{}_l Y^l{}_{;j}X^j + Z^k{}_l X^l{}_{;j}Y^j \end{aligned}. \tag{2.78}$$

Combining eqns (2.76), (2.77) and eqn (2.78) we obtain

$$R(X,Y)Z = (Z^i{}_{;lk} - Z^i{}_{;kl})X^k Y^l. \tag{2.79}$$

Comparing eqn (2.75) and eqn (2.79) gives

$$Z^i{}_{;lk} - Z^i{}_{;kl} = R^i{}_{jkl}Z^j, \tag{2.80}$$

which is the same as

$$\nabla_k\nabla_l Z^i - \nabla_l\nabla_k Z^i = R^i{}_{jkl}Z^j. \tag{2.81}$$

The last equation above could also be taken as the defining equation for the components of the curvature tensor. As shown by the left-hand side of eqn (2.81), the Riemann curvature tensor provides the measure of non-commutation of a tangent vector when parallel transported along different curves to arrive at the same space-time point.

In place of the vectors X,Y, and Z let us choose now the basis vectors e_is. Then,

$$\begin{aligned}\nabla_{e_j}\nabla_{e_k}e_l &= \nabla_{e_j}(\Gamma^a{}_{kl}e_a) \\ &= e_j(\Gamma^a{}_{kl})e_a + \Gamma^a{}_{kl}\Gamma^h{}_{ja}e_h.\end{aligned} \tag{2.82}$$

Consider now the definition of the components of the Riemann tensor as given by eqn (2.74). In particular, if a coordinate basis is chosen then $[e_i, e_j] = 0$ and we can write

$$R^i{}_{jkl} = \langle e^i, \nabla_{e_k}\nabla_{e_l}e_j\rangle - \langle e^i, \nabla_{e_l}\nabla_{e_k}e_j\rangle.$$

Then, using eqn (2.76) and a coordinate basis, the coordinate components of the Riemann curvature tensor can be given in terms of the coordinate components of the connection as

$$R^i{}_{jkl} = \frac{\partial\Gamma^i{}_{lj}}{\partial x^k} - \frac{\partial\Gamma^i{}_{kj}}{\partial x^l} + \Gamma^i{}_{ka}\Gamma^a{}_{lj} - \Gamma^i{}_{la}\Gamma^a{}_{kj}. \tag{2.83}$$

As pointed out in Section 2.4, given the metric tensor g on M, there exists a unique, torsion-free connection on M defined by the condition $\nabla_X g = 0$, which is equivalent to the vanishing covariant derivative of the metric tensor, that is, $g_{ij;k} = 0$. Then parallel transport of vectors preserve the scalar product defined by g and $g(V, V) = \text{const.}$ along a geodesic γ, where V is the tangent to γ. Then,

$$\begin{aligned}\nabla_X(g(Y, Z)) &= Xg(Y, Z) \\ &= \nabla_X(g_{ij}Y^iZ^j) \\ &= g(\nabla_X Y, Z) + g(\nabla_X Z, Y).\end{aligned} \tag{2.84}$$

Evaluating $Y(g(Z, X))$ and $Z(g(X, Y))$ and adding the first and subtracting the second from eqn (2.84) gives,

$$\begin{aligned}g(Z, \nabla_X Y) = \tfrac{1}{2}[&Xg(Y, Z) + Yg(Z, X) \\ &- Zg(X, Y) + g(Y, [Z, X]) + g(Z, [X, Y]) - g(X, [Y, Z])].\end{aligned} \tag{2.85}$$

Choosing the basis vectors e_i in place of the vectors X, Y, and Z in eqn (2.85) gives the connection coefficients in terms of derivatives of g_{ij} and the Lie derivatives of the basis vectors,

$$g(e_i, \nabla_{e_j}e_k) = g_{im}\Gamma^m{}_{jk} = \Gamma_{ijk}. \tag{2.86}$$

Choosing a coordinate basis with $[e_i, e_j] = 0$ gives the usual *Christoffel symbols*:

$$\Gamma_{ijk} = \frac{1}{2}\left(\frac{\partial g_{ij}}{\partial x^k} + \frac{\partial g_{ik}}{\partial x^j} - \frac{\partial g_{jk}}{\partial x^i}\right). \tag{2.87}$$

Hence it follows from eqns (2.83) and (2.87) that the Riemann tensor components are expressed in terms of the metric tensor and its second derivatives when the connection defined by the metric is used. From now on we shall always mean by the connection this unique connection defined by the metric tensor.

The expression eqn (2.83) and earlier definitions imply that the Riemann tensor has the symmetry given by

$$R^i{}_{jkl} = -R^i{}_{jlk}, \tag{2.88}$$

which is equivalent to $R^i{}_{j(kl)} = 0$. Further, the curvature tensor obeys the *cyclic identity* $R^i{}_{[jkl]} = 0$ which can be written as

$$R^i{}_{jkl} + R^i{}_{klj} + R^i{}_{ljk} = 0. \tag{2.89}$$

The covariant derivatives of the Riemann tensor satisfy the *Bianchi identities* given by $R^i{}_{j[kl;a]} = 0$ which is the same as,

$$R^i{}_{jkl;a} + R^i{}_{jla;k} + R^i{}_{jak;l} = 0. \tag{2.90}$$

A straightforward proof would involve writing down each term above explicitly and then substituting from eqn (2.83) and summing. There are certain additional symmetries valid when the connection is the one induced by the metric. In this case we have

$$\Gamma_{ijk} = g_{il}\Gamma^l{}_{jk}, \quad R_{ijkl} = g_{ia}R^a{}_{jkl}, \quad \Gamma^l{}_{jk} = g^{li}\Gamma_{ijk}. \tag{2.91}$$

The Riemann tensor R_{ijkl} defined by the metric has the symmetry

$$R_{ijkl} = -R_{jikl}, \tag{2.92}$$

which means $R_{(ij)kl} = 0$. Also, in this case the Riemann tensor is symmetric in the pairs of the first two and last two indices,

$$R_{ijkl} = R_{klij}. \tag{2.93}$$

The space-time (M, g) is said to have a *flat connection* if and only if $R^i{}_{jkl} = 0$, that is, all the components of the Riemann tensor must be vanishing. This is the necessary and sufficient condition for a vector at a point p to remain unaltered after parallel transport along an arbitrary closed curve through p. This is subject to the condition that all such curves can be shrunk to zero, in which case the space-time has to be simply connected. In general, the usual concept of parallel transport of vectors breaks down in a space-time manifold in the sense that given a connection, if we parallel transport a given vector along two different space-time curves to arrive at the same point, the resultant vector will be different in each case. However, when all the components of the Riemann tensor vanish, it can be shown that whenever a vector is transported from one point to the other in the space-time, the result is independent of the path taken. In such a case, the connection is also said to be *integrable* and a necessary and sufficient condition for this to happen is the vanishing of all the components of the Riemann tensor. When a symmetric connection is integrable, the manifold is called *flat.* Further, in the case of the connection being the metric connection, the vanishing of all the Riemann tensor components provides a necessary and sufficient condition for the space-time metric to be *flat*; that is, there exists a global coordinate system in M such that the metric reduces to the diagonal form with values ± 1 everywhere.

The *Ricci tensor* is defined as a type $(0,2)$ tensor which is obtained by contracting the Riemann tensor in the following manner

$$R_{jl} = R^{i}{}_{jil}. \tag{2.94}$$

As a consequence of symmetry properties discussed above, it follows that the Ricci tensor is symmetric, and also the following holds

$$R^{i}{}_{ikl} = 0. \tag{2.95}$$

A further contraction of the Ricci tensor gives the *curvature scalar* R, which is defined as

$$R = g^{ij} R_{ij}. \tag{2.96}$$

The quantity R has the property that it depends only on g_{ij} and on their derivatives only up to the second order. Further, it is linear in the second derivatives of the metric components. The total number of independent scalars that could be constructed from the metric and its derivatives up to second order is 14.

As a consequence of various symmetries listed above, the total number of independent components of R_{ijkl} reduces to 20 when the dimension of the manifold is chosen to be four. For example, when the dimension is three, R_{ijkl} has six independent components essentially given by R_{ij}, and when the dimension is two there is only one independent component, which is essentially R.

Another important tensor one could construct from R_{ijkl} is the *Weyl tensor*, which is also sometimes called the *Weyl conformal tensor*, given as:

$$C_{ijkl} = R_{ijkl} + \left\{ g_{i[l} R_{k]j} + g_{j[k} R_{l]i} \right\} + \tfrac{1}{3} R g_{i[k} g_{l]j}. \tag{2.97}$$

The symmetry properties of the Weyl tensor follow from the symmetries of the Riemann curvature tensor discussed above in that it possesses the same symmetries as the Riemann tensor. Additionally, it can be verified that the following identically vanishes

$$g^{ik} C_{ijkl} = 0. \tag{2.98}$$

The Weyl tensor is that part of the curvature tensor for which all contractions vanish for any pair of contracted indices,

$$C^{i}{}_{jil} = 0. \tag{2.99}$$

If the Weyl tensor vanishes throughout the space-time, that is, $C_{ijkl} = 0$ at all points, then one could show that the metric g_{ij} must be *conformally flat.* This means that there exists a conformal function $\Omega(x^i)$, $0 < \Omega < \infty$, such that one could write

$$g_{ij} = \Omega^2 \eta_{ij},$$

where η_{ij} is the flat Minkowskian metric. In fact, the Weyl tensor is conformally invariant in the sense that under a conformal transformation $g_{ij} \to \overline{g}_{ij} = \Omega^2 g_{ij}$ we have

$$\overline{C}^{i}{}_{jkl} = C^{i}{}_{jkl}. \tag{2.100}$$

It is possible to show that a necessary and sufficient condition for the space-time metric to be conformally flat is that the Weyl tensor must vanish everywhere.

We finally derive here the *geodesic deviation equation*, which is also called the *Jacobi equation.* This characterizes the coming together or moving away of space-time geodesics from each other as a result of the space-time curvature. Consider a smooth one-parameter family of affinely parametrized non-spacelike geodesics, characterized by the parameters (t, v), where t is the affine parameter along a geodesic and $v =$ const. characterizes different geodesics in the family with $t, v \in \mathbb{R}$ (Fig. 9).

Such non-spacelike geodesics span a two-dimensional submanifold on which t and v could be chosen as coordinates. The vectors $T = \partial/\partial t$, and $V = \partial/\partial v$ are then coordinate vectors for which $[T, V] = 0$. Then, since the torsion tensor is vanishing, we have

$$\nabla_T V = \nabla_V T$$

which implies $T^i \nabla_i V = V^i \nabla_i T$. Further, T being tangent to the geodesics, $T^i \nabla_i T^j = 0$. Now, define the operator D by $D \equiv T^i \nabla_i$. Then,

$$DV^j = V^i \nabla_i T^j.$$

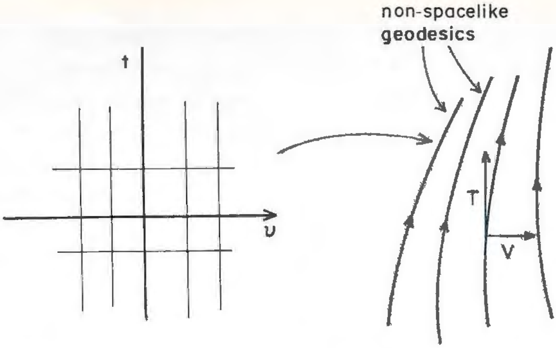

Fig. 9 A one-parameter family of non-spacelike geodesics with the tangent vector T and separation vector V.

Taking another derivative,

$$\begin{aligned} D^2V^j &= DV^i\nabla_i T^j + V^i D(\nabla_i T^j) \\ &= (T^k\nabla_k V^i)(\nabla_i T^j) + V^i T^l \nabla_l \nabla_i T^j. \end{aligned} \tag{2.101}$$

However, by the definition of the Riemann curvature tensor,

$$\nabla_l\nabla_i T^j - \nabla_i\nabla_l T^j = R^j{}_{kli}T^k.$$

Substituting this into eqn (2.101),

$$\begin{aligned} D^2V^j &= (V^k\nabla_k T^i)(\nabla_i T^j) + \nabla_i\nabla_l T^j V^i T^l + R^j{}_{kli}T^k V^i T^l \\ &= V^k((\nabla_k T^i)(\nabla_i T^j) + (\nabla_k\nabla_l T^j)T^l) + R^j{}_{kli}T^k V^i T^l \\ &= V^k(\nabla_k(T^i\nabla_i T^j)) + R^j{}_{kli}T^k V^i T^l \\ &= R^j{}_{kli}T^k V^i T^l. \end{aligned} \tag{2.102}$$

The equation

$$D^2V^j = -R^j{}_{kil}T^k V^i T^l \tag{2.103}$$

is called the Jacobi equation or the equation of geodesic deviation. It is clear from the above that $D^2V^j = 0$ if and only if all the components of the Riemann tensor are vanishing. On the other hand, whenever some components of the same are non-zero, then the neighbouring non-spacelike geodesics will necessarily accelerate towards or away from each other.

2.8 Einstein equations

We discuss in this section the Einstein equations on a space-time manifold. Throughout our discussion in this book the space-time is modelled by a pair (M, g), where M is a four-dimensional differentiable manifold and g is a Lorentzian metric tensor with the properties discussed in Section 2.5. Further, M will be assumed to have reasonable topological properties as described in Section 2.3, such as paracompactness, connectedness, Hausdorff nature, and so on. We have been referring to such a model as the *space-time manifold.* The Einstein equations to be discussed here involve the second derivatives of the metric tensor. Thus, we assume that the metric components are at least C^2 functions of the coordinates. All pairs (M', g') which are diffeomorphic to (M, g) are regarded as equivalent and we study (M, g) which represents this entire equivalence class of space-times with equivalent physical properties.

The principle of local causality and the local conservation of the energy and momentum will be accepted as the basic physical postulates for the space-time manifold (see for example, Hawking and Ellis, 1973). The basic criterion accepted by Einstein while formulating the general theory of relativity was that it is the matter distribution which determines the geometry of the space-time in terms of the Riemann curvature tensor. Next, the motion of any test particle in such a gravitational field is always independent of its own mass and composition. This is the principle of equivalence, which has been verified now to a great degree of accuracy to show that any two objects with different masses and different compositions always arrive at the same time on the surface of the earth when left from the same height. A logical consequence of this fact is that any frame of reference uniformly accelerated with respect to an inertial frame of the special theory of relativity is locally identical to a frame at rest in a gravitational field. Finally, in general relativity, one postulates the *principle of general covariance*, namely that all the physical laws are expressed as tensor equations so that they are valid in a general frame of reference and are invariant under arbitrary coordinate transformations. When restricted to the frame of free fall, these must produce the laws of special relativistic physics.

There are matter fields defined on a space-time such as an electromagnetic field or dust. All such physical fields will be assumed to be represented by a second rank tensor T^{ij}, called the *energy momentum tensor*, in the sense that T^{ij} would vanish on any open region in the space-time if and only if all the matter fields vanished on that region. Such matter fields then obey tensor equations on the space-time and the derivatives involved will be only the covariant derivatives defined with respect to the unique connection defined by the metric tensor. This is because, for any other connection defined on M, its difference with the metric connection, which is a tensor again as shown in Section 2.4, could always be regarded as another physical field on M. Such a stress energy tensor T^{ij} then describes all matter fields such as an electromagnetic field, a scalar field, or a perfect fluid. For example, in the case of *dust*, which is the matter distribution composed of non-interacting material particles, the field is characterized by the proper density ρ_0 of the flow and the four velocities of the particles given by $dx^i/d\tau$, where τ is the proper time along the timelike trajectory describing the particle worldline. The simplest second rank tensor constructed from these two quantities is given as

$$T^{ij} = \rho_0 u^i u^j$$

The component T^{00} of this energy momentum tensor is given by

$$T^{00} = \rho_0 \frac{dx^0}{d\tau}\frac{dx^0}{d\tau}.$$

In a special relativistic frame of reference, this can be interpreted as the relativistic energy density of the matter. One can further show that requiring this tensor to have zero divergence in such a frame implies the conservation of the energy as well as momentum.

Next, a *perfect fluid* is characterized by an additional scalar quantity, which is the pressure $p = p(x^i)$. In the limit as the pressure vanishes, this must reduce to the dust form of matter. Further, one also demands the conservation laws in a special relativistic frame, and that these should reduce to the classical equations of continuity and the Navier–Stokes equations in the appropriate limits. Then this energy-momentum tensor is written in a general frame as

$$T^{ij} = (\rho + p)u^i u^j + pg^{ij}$$

which could be taken as the definition of a perfect fluid in the general theory of relativity. In general, one could construct the energy-momentum tensors of various fields by using a variational principle where one has a proposed Lagrangian and the change in action is considered due to the change in the metric.

For an arbitrary frame and for other matter fields such as the electromagnetic field, or a charged scalar field, the principle of local conservation of energy and momentum states that

$$T^{ij}{}_{;j} = 0. \tag{2.104}$$

The equation above for the stress-energy tensor contains considerable information on the matter fields in a space-time. For example, if the space-time contains a Killing vector K^i then the above could be integrated to give a conservation law. The conserved vector in such a case is defined as $P^i = T^{ij}K_j$ and we get $P^i{}_{;i} = 0$ as a consequence of eqn (2.104) and the Killing equation $K_{i;j} + K_{j;i} = 0$. Then the integration of $P^i{}_{;i}$ over a compact region implies that the total flux over a closed surface of the energy-momentum is zero in the direction of the Killing vector (see for example, Hawking and Ellis,1973). Even when the space-time does not admit a Killing vector, given any point p one could set up a Riemannian normal coordinate system at p so that the metric components have the Minkowskian values and the connection coefficients $\Gamma^i{}_{jk}$ vanish at p. One could then choose a small enough neighbourhood of p so that the values of g_{ij} and $\Gamma^i{}_{jk}$ differ by an arbitrarily small amount from values at p. Using this fact it could be shown that isolated test particles should move along timelike geodesics (Fock, 1939; Dixon, 1970).

Further, all matter fields are supposed to obey the postulate of *local causality*, which is given by the statement that the equations governing the matter fields are such that given any $p \in M$, there is an open neighbourhood U of p in which a signal can be sent between any two points of U if and only if there exists a non-spacelike curve between these points. This principle is valid in the special theory of relativity and is also accepted in general relativity. The general theory of relativity is a theory of gravitation defined on a space-time manifold where the force of gravity is described in terms of the space-time curvature. These curvatures are in turn generated by the matter fields, as governed by the Einstein equations which we discuss in this section.

The above principles effectively imply that it is the space-time metric, and the quantities derived from it, that must appear in the equations for physical quantities and that these equations must reduce to the flat space-time case when the metric is Minkowskian. This is the basic content of the general theory of relativity where the space-time manifold is now allowed to have topologies other than $\mathbb{R}^4$ and the metric g_{ij} could be non-flat. In general relativity, the matter fields expressed by the stress-energy tensor are related to the non-flat nature of space-time by means of the Einstein equations, which are the basic equations satisfied by the space-time metric. In Einstein's theory, one does not discuss the physical interaction of matter fields in a fixed background metric prescribed in advance. Actually, g_{ij}s are treated as dynamical variables which depend on the matter content of the space-time and are to be solved from the Einstein equations.

An important indicator towards obtaining this relationship between the matter content and space-time geometry is provided by the Newtonian theory where the gravitational field is described by a potential ϕ. The tidal acceleration between nearby particles is given in terms of the separation between them and second derivatives of ϕ. In a curved space-time manifold, such tidal accelerations are described by the Jacobi eqn (2.103) in terms of the Riemann curvature tensor. Further, we must recover the Poisson equation

$$\nabla^2\phi = 4\pi\rho, \tag{2.105}$$

in the Newtonian limit. Now, both in the special and general theory of relativity the matter content is described by the stress-energy tensor T_{ij} and the mass-energy density ρ corresponds to the quantity $T_{ij}V^iV^j$. Thus, each side of the Poisson's equation corresponds to the Riemann tensor as expressed in the Jacobi equation and $T_{ij}V^iV^j$ respectively. Another important indicator for this comparison is provided by the Bianchi identities (2.90). Contracting i with k in eqn (2.90) gives

$$\nabla_a R_{jl} + \nabla_i R^i{}_{jla} - \nabla_l R_{ja} = 0.$$

Next, raising j and contracting with a,

$$\nabla_a R^a{}_l + \nabla_i R^i{}_l - \nabla_l R = 0$$

which is the same as $2\nabla_i R^i{}_l - \nabla_l R = 0$. Thus, we get

$$\nabla^i G_{ij} = 0, \tag{2.106}$$

where G_{il} is the *Einstein tensor* defined by

$$G_{il} \equiv R_{il} - \tfrac{1}{2}g_{il}R.$$

One could try, as a result of the indications given above, to make a a direct comparison given by $R_{ij} = 4\pi T_{ij}$ as field equations in general relativity. However, this does not work because the contracted Bianchi identities imply $\nabla_l R = 0$ and so the trace $T = \text{const.}$ throughout the space-time. This is an unphysical restriction on the matter content.

As a result of the indicators above, Einstein proposed the field equations

$$G_{ij} = R_{ij} - \tfrac{1}{2}Rg_{ij} = 8\pi T_{ij}. \tag{2.107}$$

In this case, in fact the contracted Bianchi identities imply the local conservation of energy and momentum through the Einstein equations. Taking the trace of the eqns (2.107) we get

$$R = -8\pi T.$$

Substituting this back in eqn (2.107) gives the alternative form of the Einstein equations

$$R_{ij} = 8\pi(T_{ij} - \tfrac{1}{2}Tg_{ij}). \tag{2.108}$$

It is clear by considering the definition of the Ricci tensor that the Einstein equations depend on the derivatives of g_{ij} up to the second order and that they are highly non-linear in g_{ij}s. It may be noted, however, that the Einstein equations are linear in the second derivatives of g_{ij}. In fact, the quantities R_{ij} and Rg_{ij} are the only second-rank symmetric tensors which are linear in the second derivatives of the metric and involve only up to second derivatives of g_{ij}. Actually, the Einstein equations are a coupled system of non-linear second order partial differential equations for g_{ij}s. This makes the task of solving them extremely difficult. One generally needs to impose several symmetry assumptions on the space-time in order to work out the metric components as a solution to the Einstein equations. Several solutions to the Einstein equations will be discussed in the next chapter; these are used in the later chapters and are physically important.

Given the energy momentum tensor T^{ij}, the field equations may be viewed as the set of differential equations to determine the gravitational potentials g^{ij} to determine the resulting geometry. A particularly important case here is that of vacuum solutions when $T^{ij} = 0$. On the other hand, one could arbitrarily specify the ten metric potentials, then the one could compute the Einstein tensor G_{ij} and then the field equations determine the energy-momentum tensor T_{ij}. However, in this case, the resulting T^{ij} turns out to be unphysical most of the time in that it may violate the energy conditions ensuring the positivity of mass-energy density. Such a violation of the energy conditions is rejected on physical grounds in that all observed classical fields obey such a positivity of energy density, which is closely connected with the physical features of gravitation theory.

In general, the field equations are ten equations connecting the total of twenty quantities which are ten components g_{ij} and the other ten components of T_{ij}. Thus, the field equations are the conditions placing constraints on the simultaneous choice of these twenty quantities. If part of the gravitational potentials and the matter contents are determined from physical conditions, then such conditions are used to determine the matter and geometry fully. In particular, if one considers the vacuum equations

$$G_{ij} = R_{ij} - \tfrac{1}{2}Rg_{ij} = 0,$$

then there are ten equations to determine the ten quantities g_{ij}. However, the Bianchi identities

$$\nabla_j G^{ij} = 0$$

place four differential constraints on these equations which are not all independent. Thus, there is an indeterminancy in that there are fewer equations as compared to the unknowns to be determined. Further, there is an intrinsic gauge freedom available in the general theory of relativity which does not allow a complete determination of the metric potentials. This is given by the coordinate freedom which allows a transformation from one set of coordinates x^i to any other set of coordinates $x^{i'}$. One could, however, use this coordinate freedom to impose conditions on the metric components. For example, choosing the normal coordinates gives $g_{00} = 1$ and $g_{0\alpha} = 0, \alpha = 1, 2, 3$ in this coordinate system. This leaves six other components to be determined from the field equations. This issue is closely connected with the Cauchy problem in general relativity where the basic problem is, given an initial data on a regular spacelike hypersurface one would like to determine its unique evolution in the future or past.

Finally, we discuss here the Einstein equations with a cosmological term. It may be noted that the most general second rank tensor which can be constructed out of G_{ij} and g_{ij} so that it is divergence free and involves the derivatives of the metric tensor up to second order only is the linear combination $G_{ij} + \Lambda g_{ij}$ (Lovelock, 1972), where Λ is a constant. Thus, addition of such a constant multiple of g_{ij} to the Einstein tensor preserves all the required properties of equations (2.107) discussed above. Einstein historically introduced the *cosmological term* Λ in his equations in order to generate static cosmological solutions, and wrote the equations as

$$R_{ij} - \tfrac{1}{2} R g_{ij} + \Lambda g_{ij} = 8\pi T_{ij}. \tag{2.109}$$

If $\Lambda \neq 0$ then one does not obtain the Newtonian theory in the limit of slow motions and weak fields; however, if the magnitude of Λ is very small then such departures will be quite negligible and approximate agreement with the Newtonian theory is obtained. We shall discuss the issue of limits on the values of Λ in Chapter 8. It is seen that for an empty space-time with $T_{ij} = 0$, the Einstein equations simplify to

$$R_{ij} = \Lambda g_{ij}.$$

3

SOLUTIONS TO THE EINSTEIN EQUATIONS

The purpose of this chapter is to discuss several useful space-times which are exact solutions of Einstein equations, and to review some of their properties that are referred to in later chapters.

The Einstein equations form a highly non-linear system of differential equations and due to their complexity, a completely general solution is not known. Thus, the known exact solutions usually assume a rather high degree of symmetry such as the spherical or axial symmetry, and the existence of necessary Killing vector fields on the space-time, and to that extent represent an idealized situation. However, such space-time examples provide a good idea of what is possible within the framework of the general theory of relativity outlined in the previous chapter. They also give an indication of certain important features that may be present in a situation which is more general than the actual solution being studied. Thus, some of such solutions have significant physical applications. For example, the Minkowski space-time is both the geometry of the special theory of relativity and locally that of any general relativistic model. Again, the spherically symmetric and asymptotically flat space-times such as the Schwarzschild and Kerr geometries are useful to model the space-time outside the sun and stars and could be used to obtain conclusions relevant for the experimental verification of the general theory of relativity. Such solutions could also possibly represent the outcome of a complete gravitational collapse of a massive star. The other models discussed here also have interesting implications, particularly on the issue of the final fate of gravitational collapse and cosmology.

A large number of exact solutions to the Einstein equations are known which are obtained under various symmetry conditions and studied mostly locally. The examples discussed here aim at either studying certain global properties of interest or reviewing results which would be needed for a later chapter. Section 3.1 studies the Minkowski space-time and its conformal compactification. The Schwarzschild space-time and Kruskal–Szekeres extension are discussed in Section 3.2 and non-spacelike trajectories in Kerr–Newman metrics are reviewed in Section 3.3. The charged versions of the Schwarzschild and Kerr space-times are discussed in Section 3.4. The Vaidya radiating metric is considered in Section 3.5 and the final section studies the Robertson–Walker cosmological models.

3.1 Minkowski space-time

The Minkowski space-time is mathematically the manifold $M = \mathbb{R}^4$ with the Lorentzian metric

$$ds^2 = -dt^2 + dx^2 + dy^2 + dz^2, \tag{3.1}$$

with $-\infty < t, x, y, z < \infty$ giving the range of the coordinates. This is a flat space-time with all the components of the Riemann tensor $R^i_{jkl} = 0$, and hence the simplest empty space-time solution to Einstein equations

$$G_{ij} = 8\pi T_{ij} = 0,$$

which underlies the physics of special theory of relativity. The vector $\partial/\partial t$ provides a time orientation for this model. If we use the spherical polar coordinates (t, r, θ, ϕ) given by $x = r\sin\theta\sin\phi, y = r\sin\theta\cos\phi$, and $z = r\cos\theta$ then eqn (3.1) becomes

$$ds^2 = -dt^2 + dr^2 + r^2(d\theta^2 + \sin^2\theta\, d\phi^2). \tag{3.2}$$

The range of the coordinates r, θ, ϕ is $0 < r < \infty, 0 < \theta < \pi$ and $0 < \phi < 2\pi$. Two such coordinate neighbourhoods are needed to cover all of the Minkowski space-time.

As discussed in Chapter 2, all the components of the Riemann curvature tensor vanish for the Minkowski space-time which is a flat space-time. In the Minkowskian coordinate system (t, x, y, z) this is obvious as all the metric components are constants and hence all the connection coefficients will vanish. In other coordinate systems such as a spherical system (t, r, θ, ϕ), the connection coefficients Γ^i_{jk} will not all vanish (for example, $\Gamma^1_{22} = r$); however, all the Riemann curvature components will still be vanishing.

The *Lorentz transformations* on the Minkowski space-time are defined as the set of those metric preserving isometries which are linear and homogeneous transformations. Physically, these represent the change of reference frame from one inertial observer to an another inertial observer. Thus, the Lorentz transformations are defined by the coordinate change

$$x^i \to x^{i'} = L^i{}_j x^j.$$

From the above and the fact that these are metric preserving isometries, it follows that $\det L^i{}_j = \pm 1$ and hence the matrix is non-singular. If $\det L^i{}_j = +1$ and further $L^0_0 \geq 1$, then the Lorentz transformation preserves both the orientations in space as well as time. The set of all the Lorentz transformations form a group where the identity map is given by δ^i_j and the inverse is defined by the inverse matrix. The Lorentz group is a subgroup of the Poincaré group of transformations which are general inhomogeneous mappings that leave the Minkowskian metric invariant. Such a general mapping consists of a Lorentz transformation together with an arbitrary translation in space and time. This is a ten-parameter group, consisting of six Lorentz parameters and four translation parameters, and represents physically a mapping of one inertial frame into another in a general position in the space and time.

The geodesics of Minkowski space-time are the straight lines of the underlying Euclidian geometry. Given an event in M, the lines at 45° to the time axis through that event give null geodesics in M. Such null geodesics form the boundary of the chronological future or past $I^{\pm}(p)$ of an event p which contains all possible timelike material particle trajectories through p including timelike geodesics. The causal future $J^{+}(p)$ is the closure of $I^{+}(p)$ in Minkowski space, which includes all the events in M which are either timelike or null related to p by means of future directed non-spacelike curves from p. The family of spacelike hypersurfaces $t =$ const. in the Minkowski space-time gives a family of Cauchy surfaces which covers all of M. (A Cauchy surface is a spacelike hypersurface in the space-time such that all inextendible non-spacelike curves in M meet this surface once and only once. We discuss this concept further in Section 4.4. However, all spacelike hypersurfaces in M need not be Cauchy surfaces. For example, the family given by

$$-t^2 + x^2 + y^2 + z^2 = A = \text{const.}$$

with $A < 0$ are inextendible spacelike surfaces which are not Cauchy surfaces. All these surfaces are fully contained inside the chronological past or the future of the origin and there are timelike geodesics outside this past or future cone which do not meet any of these surfaces.

To understand the global properties and structure of infinity of the Minkowski space-time, we can use the procedure given by Penrose (1968) and Geroch, Kronheimer and Penrose (1972). An arbitrary event p in the Minkowski space-time is uniquely determined either by its chronological future $I^{+}(p)$ or past $I^{-}(p)$. If a future directed non-spacelike curve γ has a future end point at p, we have $I^{-}(\gamma) = I^{-}(p)$. (By definition, $I^{-}(\gamma)$ is the union of all $I^{-}(q)$ with q being a point on the curve γ.) On the other hand, if γ is future inextendible without any future end point, the set $I^{-}(\gamma)$ determines a 'point at infinity' of M. (A future or past inextendible curve, in the context of Minkowski space-time, is a trajectory which goes off to the infinity in future or past without stopping anywhere. For a precise definition, we refer to Section 4.1.) Two such curves γ_1, and γ_2 determine the same ideal point or a point at infinity if $I^{-}(\gamma_1) = I^{-}(\gamma_2)$. Such a procedure defines future ideal points. Past ideal points are defined dually

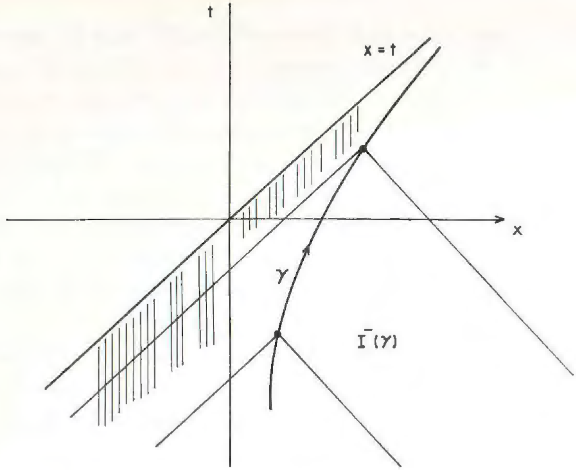

Fig. 10 The pasts of the timelike curve γ and the null hypersurface $x = t$ coincide.

using past inextendible non-spacelike curves. In the Minkowski space-time, there are future directed inextendible timelike curves γ which have the same past, which is the entire space-time M, that is, $I^-(\gamma) = M$. Hence, all such timelike curves determine a single future ideal point i^+, called the future timelike infinity. The past timelike infinity i^- is similarly defined. If we choose γ to be a future endless null geodesic or a null curve, it is possible to have a situation where $I^-(\gamma)$ is not the entire Minkowski space-time. Certain timelike curves also have this property. For example, consider the past of the timelike hyperbola

$$t = \sinh \lambda, \quad x = \cosh \lambda, \quad y, z = 0, \quad -\infty < \lambda < \infty. \tag{3.3}$$

Then, $I^-(\gamma)$ lies completely to the past of the null hypersurface $x = t$. It can be shown in general (see Geroch, Kronheimer and Penrose, 1972) that if for a non-spacelike curve γ, if $I^-(\gamma) \neq M$ and if γ is future endless, then there exists a null hypersurface S_γ, the half-space below which coincides with $I^-(\gamma)$ (Fig. 10).

If we denote the collection of ideal points so defined by $\mathcal{I}^+$, then there is a one-one and onto correspondence between the points of $\mathcal{I}^+$ and such null hypersurfaces. Any such null hypersurface is determined by the value of the time t at which it intersects the time axis and by the direction of null vector at the point of intersection. Since the set of all possible light rays directions at any point is equivalent to the two-sphere S^2, it follows that $\mathcal{I}^+$ is a three-dimensional manifold with topology $S^2 \times \mathbb{R}$.

The three-dimensional null hypersurfaces $\mathcal{I}^+$ and $\mathcal{I}^-$ are called the future and past null infinities respectively for the Minkowski space-time. As we shall discuss in Chapter 4, a general space-time also would admit such a boundary construction under certain conditions such as being asymptotically flat and empty. One can show for the Minkowski space-time that all complete null hypersurfaces are flat and so are like the surfaces $\{x = t\}$, in which case the topological structure of the null infinity is clearly $\mathcal{I}^+ = S^2 \times \mathbb{R}$. It is not clear, however, that the null infinities will necessarily have the same topological structure even in the case of a general space-time, which we discuss in Section 4.7.

It is possible to introduce a differential structure as well as a metric on $\mathcal{I}^+$. To see this, we first note that a convenient way to attach the ideal point boundary $\mathcal{I}^+$ to M is to use a suitable conformal factor Ω to obtain a transform of the original space-time metric η_{ij},

$$g_{ij} = \Omega^2 \eta_{ij}, \quad \Omega > 0, \tag{3.4}$$

which leaves the causal structure of M invariant because the null geodesics of η_{ij} and the unphysical metric g_{ij} are the same up to a reparametrization as discussed earlier. Thus, the past of any non-spacelike curve γ is unchanged and there is a natural correspondence between ideal points in two space-times. Since light cones are unaltered by a conformal transformation, the boundary attachment obtained in this manner is coordinate independent.

In the metric (3.2), one could introduce the *advanced* and *retarded* null coordinates given by

$$v = t + r, \quad u = t - r, \tag{3.5}$$

which gives a reference frame based on null cones, which is most suitable to analyse the radiation fields (Fig. 11).

Under this transformation of coordinates, the metric becomes,

$$ds^2 = -dudv + \tfrac{1}{4}(u - v)^2(d\theta^2 + \sin^2\theta\, d\phi^2) \tag{3.6}$$

with $-\infty < v < \infty$ and $-\infty < u < \infty$. Now, the information at future null infinity corresponds to taking limit as $v \to \infty$, which amounts to moving in future along $u =$ const. light cones. Similarly, past null infinity corresponds to $u \to \infty$. This procedure could be made precise in a coordinate independent way. We can compactify the Minkowski space-time M by means of a conformal transformation of eqn (3.6) given by

$$\Omega^2 = (1 + v^2)^{-1}(1 + u^2)^{-1}, \tag{3.7}$$

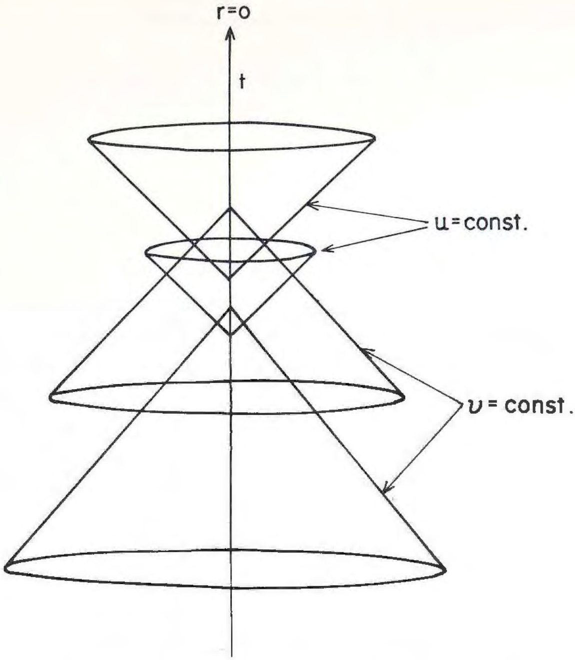

Fig. 11 In the Minkowski space-time, the future light cones are given as the null surfaces u = const. Similarly, the past light cones are given as v = const.

and then by adding the closure to add the null infinities. We also introduce new coordinates p, q by

$$v = \tan p, \quad u = \tan q. \tag{3.8}$$

Then, the corresponding ranges for p and q are

$$-\frac{\pi}{2} < p < \frac{\pi}{2}, \qquad -\frac{\pi}{2} < q < \frac{\pi}{2}$$

and the metric $\bar{g}_{ij}$ on the unphysical space-time $\overline{M}$, after the conformal transformation, is given by

$$d\bar{s}^2 = -dpdq + \sin^2(p-q)(d\theta^2 + \sin^2\theta \, d\phi^2). \tag{3.9}$$

It is possible to see now that the metric (3.9), with the coordinate ranges given above, is a manifold embedded as a part of the Einstein static universe. To see this, write

$$T = p + q, \quad R = p - q, \tag{3.10}$$

then eqn (3.9) becomes in (T, R, θ, ϕ) coordinates

$$d\bar{s}^2 = -dT^2 + dR^2 + \sin^2 R(d\theta^2 + sin^2\theta \, d\phi^2), \tag{3.11}$$

with the coordinate ranges

$$-\pi < T + R < \pi, \qquad -\pi < T - R < \pi. \tag{3.12}$$

This is precisely the natural Lorentz metric on $S^3 \times \mathbb{R}$, which is the Einstein static universe, except that the coordinate ranges are now restricted by eqn (3.12). In this picture, the future null infinity $\mathcal{I}^+$ is given by $T = \pi - R$ for $0 < R < \pi$ and the past null infinity is given by $T = -\pi + R$ for $0 < R < \pi$.

This conformal structure of infinity for the Minkowski space-time is shown in Fig. 12. In fact, the structure of infinity for any spherically symmetric space-time can be depicted by a similar diagram which is called a *Penrose diagram*. As mentioned above, $\mathcal{I}^+$ is topologically $S^2 \times \mathbb{R}$ and this discussion on Minkowski space-time will be useful when we define general asymptotically flat space-times in Chapter 4.

Finally, in order to have a better insight into the asymptotic structure of the Minkowski space-time (and hence of general asymptotically flat space-times to be discussed in Chapter 4), we work out below the light cone cuts of future null infinity for the Minkowski space-time (Joshi, Kozameh and Newman, 1983). The light cone evolves from an arbitrary apex in the space-time to future null infinity and its intersection with $\mathcal{I}^+$ is obtained. Both the light cone and $\mathcal{I}^+$ are three-dimensional null hypersurfaces in M and hence their intersection is a two-surface at $\mathcal{I}^+$. It will be shown in Chapter 4 that the knowledge of such cuts yields considerable information about the interior space-time and the metric in the neighbourhood of the apex point.

First we introduce a coordinate system on M which is more suitable for this purpose. Using the retarded time u, and a complex stereographic coordinate ζ and its complex conjugate $\bar{\zeta}$ on the sphere defined by

$$\zeta = e^{i\phi} \cot(\theta/2), \tag{3.13}$$

the Minkowski metric (3.2), when written in the $(u, r, \zeta, \bar{\zeta})$ system is given by

$$ds^2 = -du^2 - 2du dr + r^2 \frac{d\zeta d\bar{\zeta}}{P_0^2}, \tag{3.14}$$

where $P_0 = (1 + \zeta\bar{\zeta})/2$. Following Exton, Newman and Penrose (1969) we define $u' = (1/\sqrt{2})u$ and $r' = \sqrt{2}r$, which are more convenient variables for the study of asymptotic structure. Since null cones and null geodesics

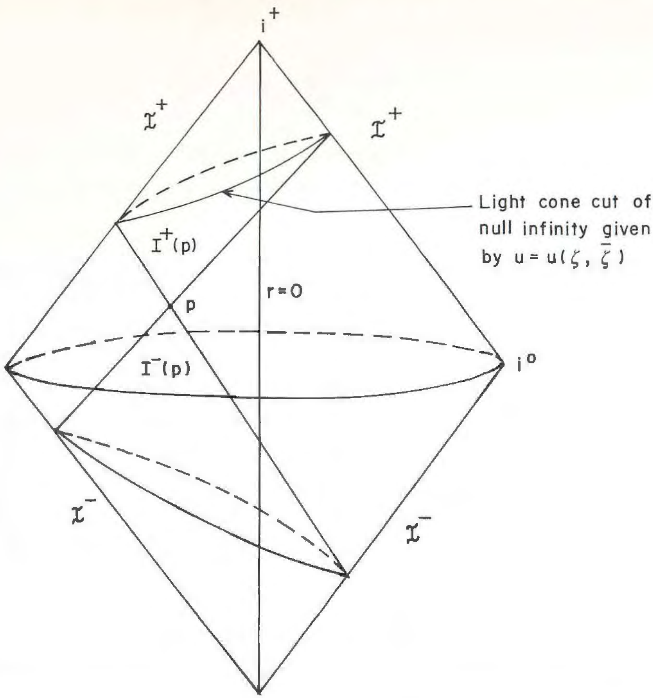

Fig. 12 Conformal infinity in the Minkowski space-time. The future and past null infinities $\mathcal{I}^{\pm}$ are both topologically $S^2 \times \mathbb{R}$. Every null geodesic in the space-time meets $\mathcal{I}^+$ in future and $\mathcal{I}^-$ in the past. Here the point i^+ denotes the future timelike infinity, and past timelike infinity is similarly defined where every timelike geodesic terminates in future and past respectively. The spacelike infinity is denoted by i^o. The light cone cuts of null infinity for an event p are shown.

are conformally invariant, there is an inherent conformal freedom available. Suppressing the primes, we conformally transform eqn (3.14) to

$$d\bar{s}^2 = \Omega^2 ds^2 = -4\ell^2 du^2 + 4du d\ell + \frac{d\zeta d\bar{\zeta}}{P_0^2}, \tag{3.15}$$

where we have introduced a new variable $\ell = 1/\sqrt{2}r$ and the conformal factor is chosen to be $\Omega = \sqrt{2}\ell$. In the limit as $r \to \infty$, we have $\ell \to 0$ and the null infinity $\mathcal{I}^+$ is defined by the condition $\ell = 0$. Future directed null cones are characterized by the values of u, ζ, and $\bar{\zeta}$ and so the coordinates $(u, \zeta, \bar{\zeta})$ can be used as coordinates on $\mathcal{I}^+$, which are called the Bondi coordinates on $\mathcal{I}^+$. In these coordinates, a hypersurface of $\mathcal{I}^+$ has the metric

$$ds^2 = \frac{1}{P_0^2} d\zeta d\bar{\zeta}. \tag{3.16}$$

Thus, $\mathcal{I}^+$ is a null hypersurface which is generated by the null curves $\zeta, \bar{\zeta} =$ const. For the manifold $(\overline{M}, \bar{g})$ given by the conformal compactification as above, it is easy to see that

$$\frac{\partial \Omega}{\partial x^i} = (0, 1, 0, 0), \tag{3.17}$$

and

$$\bar{g}^{ij} \frac{\partial \Omega}{\partial x^i} \frac{\partial \Omega}{\partial x^j}_{|\mathcal{I}^+} = 0. \tag{3.18}$$

Thus, Ω is differentiable on $\overline{M}$, the new unphysical manifold with boundary, and $\partial\Omega/\partial x^i$ is a null vector. The factor Ω is smooth everywhere and $\Omega = 0$ on $\mathcal{I}^+$ which is a null hypersurface. Similarly we could discuss the past null infinity $\mathcal{I}^-$, which has a similar structure.

Next, we work out the complete light cone at any given apex point in the space-time. The null geodesic equations of the space-time are given by (with the dot denoting derivative with respect to s, the affine parameter)

$$\begin{aligned} 2\ell^2 \dot{u} - \dot{\ell} &= 1, \\ \ddot{u} + \ell \dot{u}^2 &= 0, \\ \ddot{\zeta}(1 + \zeta\bar{\zeta}) - 2\bar{\zeta}\dot{\zeta}^2 &= 0, \\ \ddot{\bar{\zeta}}(1 + \zeta\bar{\zeta}) - 2\dot{\zeta}\dot{\bar{\zeta}}^2 &= 0, \\ 4\ell^2 \dot{u}^2 - 4\dot{u}\dot{\ell} - \frac{\dot{\zeta}\dot{\bar{\zeta}}}{P_0^2} &= 0, \end{aligned} \tag{3.19}$$

where the last equation corresponds to $ds^2 = 0$. Restricting ourselves presently to the equatorial plane $\theta = \pi/2$ for the sake of simplicity, eqn (3.19) can be written as

$$\begin{aligned} 2\ell^2 \dot{u} - \dot{\ell} &= 1, \\ \ddot{u} + \ell \dot{u}^2 &= 0, \\ \ddot{\phi} = 0, \quad \dot{\phi} &= b, \\ \ell^2 \dot{u}^2 - \dot{u}\dot{\ell} = \dot{\phi}^2 &= b^2/4. \end{aligned} \tag{3.20}$$

We shall now eliminate the parameter s from the above equations. For that, one could use the first and last equations in the above to obtain $\dot{\ell}^2 = 1 - b^2\ell^2$. Thus we have

$$\dot{\ell} = \pm\sqrt{(1 - b^2\ell^2)},$$
$$ds = \pm\frac{d\ell}{\sqrt{1 - b^2\ell^2}}. \tag{3.21}$$

We note that $\dot{\ell} < 0$ corresponds to a null ray moving away from the origin. Next, if $\dot{\ell} > 0$ initially, then the ray moves initially towards the origin of the coordinate system $(r = 0, \ell = \infty)$ and after reaching a minimum value r_m, where $\dot{\ell} = \sqrt{1 - b^2\ell_m^2} = 0$, it begins to move outwards and again $\dot{\ell} < 0$. For the sake of definiteness we shall choose here rays such that initially $\dot{\ell} < 0$; however, by considering the other sheet $\dot{\ell} > 0$ as well, we can span the full light cone of null rays from our starting point.

The above equations can be written as

$$du = -\frac{d\ell}{2\ell^2\sqrt{1 - b^2\ell^2}} + \frac{d\ell}{2\ell^2},$$
$$d\phi = \frac{-b d\ell}{\sqrt{1 - b^2\ell^2}}. \tag{3.22}$$

Suppose now the apex of the cone has the values $\ell = \ell_0, u = u_0, \phi = \phi_0$. Then integrating the above equations from ℓ_0 to an arbitrary ℓ gives the equations for one sheet of the light cone. For the sake of simplicity we choose the apex on the $\phi = 0$ axis, that is, $\phi_0 = 0$. Now by taking the limit as the null rays escape to infinity, that is, $\ell \to 0$, the above integration provides us with a portion of the cut at infinity of the rays coming from u_0, ℓ_0, ϕ_0. When we are describing only the rays in the equatorial plane, integration of eqn (3.22) from ℓ_0 to 0 yields

$$u - u_0 = \frac{1 - \sqrt{1 - b^2\ell_0^2}}{2\ell_0}, \tag{3.23}$$

$$\phi = \sin^{-1}(b\ell_0). \tag{3.24}$$

The initial direction b ranges from 0 to ℓ_0^{-1}. Note that for a fixed apex one has a one-to-one relation between the initial direction b and the final angular position ϕ on the future null infinity. By eliminating b from eqns (3.23) and (3.24), we obtain the equatorial plane portion of the light cone cut which is given by

$$u = u_0 + (1/2\ell_0)(1 - \cos\phi). \tag{3.25}$$

For the sheet $\dot{\ell} < 0$ which we have been considering, $\cos\phi$ will be positive because ϕ will always be in the first or fourth quadrant in this situation. For the other sheet corresponding to $\dot{\ell} > 0$ initially, $\cos\phi$ will be negative.

The portion of the light cone cut given by eqn (3.25) describes, as we have mentioned, only an S^1 worth of null rays intersecting $\mathcal{I}^+$ since we have restricted ourselves to the equatorial plane. However, because of spherical symmetry the full cut, which is topologically S^2, can be generated by rotating this plane. We shall discuss such asymptotic cuts of null infinity in Chapter 4.

3.2 Schwarzschild geometry

The Schwarzschild solution represents the geometry exterior to a spherically symmetric massive body such as a star and has been used extensively to verify the predictions of the general theory of relativity experimentally. This is the empty exterior solution where the Ricci tensor vanishes and which is matched at the boundary to the interior solution inside the body. In (t, r, θ, ϕ) coordinates, the metric can be given in the form

$$ds^2 = -\left(1 - \frac{2m}{r}\right) dt^2 + \left(1 - \frac{2m}{r}\right)^{-1} dr^2 + r^2 d\Omega^2, \tag{3.26}$$

where $d\Omega^2 = d\theta^2 + \sin^2\theta \, d\phi^2$. Here the coordinate t is timelike and the other three coordinates r, θ, ϕ are spacelike. The radial coordinate r has the property that the two-sphere given by $t =$ const., $r =$ const. has the two-metric given by

$$ds^2 = r^2(d\theta^2 + \sin^2\theta \, d\phi^2).$$

It follows that the area of any such two-sphere would be $4\pi r^2$.

The coordinate r is restricted by the condition $r > 2m$ because the above metric has an apparent singularity at $r = 2m$. The coordinate t has the range $-\infty < t < \infty$. The solution above is generated by solving the vacuum Einstein field equations for a spherically symmetric space-time and the quantity m appears as the constant of integration. The value of this constant can be determined by considering the weak field Newtonian limit of general relativity. If ϕ is the Newtonian gravitational potential, then in non-relativistic units,

$$g_{00} \simeq 1 + \frac{2\phi}{c^2} = 1 - \frac{2GM}{c^2 r},$$

where G is the Newtonian constant of gravity, c is the velocity of light, and M is the point mass at the origin which gives rise to the Newtonian potential ϕ. This determines the constant of integration m in the Schwarzschild solution as

$$m = \frac{GM}{c^2}.$$

Thus, the Schwarzschild solution is interpreted as describing the gravitational field of a point particle with mass m (in relativistic units $G = c = 1$) situated at the center.

Apart from predicting small observable departures from the Newtonian gravity, the Schwarzschild solution of Einstein equations is important for the theory of black holes as well. As will be discussed in Chapter 6, sufficiently massive stars unable to support themselves against the pull of self gravity must undergo a complete gravitational collapse when they have exhausted their internal nuclear fuel. The final fate of a spherically symmetric homogeneous dust collapse must be a Schwarzschild configuration which contains a space-time singularity hidden within the event horizon. This gives rise to a black hole in the space-time, which is a region from which no causal signals can reach a far-away observer. This scenario forms the basis of much of the theory and applications of the modern black hole physics.

The Schwarzschild metric is static in the sense that $\partial/\partial t$ is a timelike Killing vector which is a gradient. The metric components g_{ij} here are independent of time. Also, there are no mixed terms in eqn (3.26) involving both space and time; hence there is no rotation inherent in the space-time. To make this more precise, we discuss here briefly the stationary and static solutions of the field equations briefly. A solution will be called stationary if the time does not enter explicitly in the metric potentials. In such a case, a coordinate system will exist in which the metric components will be time-independent; that is, if x^0 is the timelike coordinate,

$$\frac{\partial g_{ij}}{\partial x^0} = 0.$$

Defining a vector $X^i = \delta^i{}_0$, it is seen that the Lie derivative $L_X g_{ij} = 0$ in this coordinate system. Since this is a tensor, it follows that this Lie derivative will vanish in all coordinate systems and hence X^i is a Killing vector. On the other hand, if the space-time admits a timelike Killing vector field, then it is possible to choose a coordinate system adapted to it such that the Lie derivative $L_X g_{ij} = 0$ in this frame. Then, the metric is again stationary. Thus, a space-time is called *stationary* if and only if it admits the existence of a timelike Killing vector field.

We note that a if a space-time is stationary, that does not mean that the metric components cannot evolve in time. It is just that the time does not enter explicitly in the solution. However, the stronger requirement of staticity means that there is no time evolution of the system, which is time-symmetric about any origin of time. In such a case, one would expect that in the coordinate system adapted to the timelike Killing vector field, the metric would admit no cross terms as well such as g_{ij} with $i \neq j$. The reason is, in such a case under a time reversal $x^0 \to -x^0$, the sign of those pieces of ds^2 containing the cross terms in g_{ij} will be reversed. However, the staticity assumption means that ds^2 must remain invariant under time reversal about any origin of time. This implies that the cross terms must vanish in the expression for ds^2. Thus, a *static* space-time is characterized by the existence of a timelike Killing vector field for the space-time, and the additional requirement that in the coordinate system adapted to this vector field the metric components are time independent and no cross terms appear in the line element ds^2. Such a property of the Killing vector field is characterized by its being hypersurface orthogonal. (A vector field X^i is called *hypersurface orthogonal* if and only if $X_{[i}\nabla_j X_{k]} = 0$.) Thus, a space-time is static if and only if it admits a timelike Killing vector field which is hypersurface orthogonal. It is also possible to check directly from the form of the metric (3.26) that the metric is time symmetric under a change $t \to -t$ and is also invariant under time translations.

As pointed out above, the Schwarzschild metric (3.26) above is the solution of the vacuum Einstein equations with the assumption of spherical symmetry on the space-time. It is clear that the coordinate system (t, r, θ, ϕ) provides a frame in which the metric components are time-independent. Thus the solution is stationary. Further, defining the vector field X^i by the condition $X^i = \delta^i{}_0$ in this coordinate system gives a Killing vector field with the components

$$X_i = (1 - 2m/r, 0, 0, 0),$$

which can be seen to be hypersurface orthogonal to the family of spacelike hypersurfaces $t =$ const. Thus the solution is seen to be static. We thus have the *Birkhoff theorem* (1923), namely that a spherically symmetric vacuum solution of the Einstein equations must be necessarily static. An important implication of this theorem is that even when a spherically symmetric star undergoes pulsations or changes in shape, while maintaining the spherical symmetry, it cannot radiate any disturbances in the exterior such as gravitational waves. It is thus shown by the Birkhoff theorem that any spherically symmetric solution of Einstein equations with $R_{ij} = 0$ is necessarily the Schwarzschild solution. Hence, the Schwarzschild exterior solution can be used to describe the outside metric for several situations such as a spherically symmetric star which is either static or which undergoes radial pulsations, or a radial spherically symmetric gravitational collapse.

The spherical symmetry of the Schwarzschild space-time M is exhibited by the fact that the metric components g_{00} and g_{11} are functions of r alone and not of θ and ϕ, and as implied by the angular part of the metric. Specifically, the isometry group of M contains a subgroup which is isomorphic to the group $SO(3)$ and the orbits of this subgroup are two-dimensional spheres (see for example, Hawking and Ellis, 1973). These isometries are interpreted as rotations and thus the metric remains invariant under rotations in general for any spherically symmetric space-time. The parameter m here serves as the source of the gravitational field and setting $m = 0$ gives the flat Minkowski space-time. As pointed out above, the comparison with Newtonian theory shows that m is to be treated as the gravitational mass of the body producing the field as measured from infinity. The space-time here is asymptotically flat because as r tends to infinity we recover a flat space-time metric and the gravitational field diminishes to zero.

Generally, eqn (3.26) is taken to represent the outside metric for a star with $r > r_0$ where r_0 gives the boundary of the star. The metric inside $r < r_0$ is a different interior metric determined by the matter distribution T_{ij} inside the star and is matched at the boundary $r = r_0$ with eqn (3.26). However, in the case of a complete gravitational collapse, when all the mass collapses at $r = 0$, it is necessary to consider the metric (3.26) as an empty space-time solution for all the values of r. Clearly this metric has singularities at $r = 0$ and $r = 2m$ and hence it represents only one of the patches $0 < r < 2m$ or $2m < r < \infty$. If we confine to the manifold M given by the later range of values of r, it is necessary to determine if M is extendible; that is, if there exists a bigger space-time (M', g') with M embedded in M' and $g = g'$ on M. That this should be possible is indicated by the fact that even though the form of the above metric is singular at $r = 2m$, the curvature scalars are all well-behaved at this point and so this could be merely a singularity due to an inappropriate choice of coordinates. A decision on whether a given space-time manifold is maximal or not can be made by looking at the geodesics of the space-time. In a maximal manifold, one would require all the geodesics to be extended in both the directions to an infinite value of their affine parameter, or they must terminate at an intrinsic singularity of the space-time which is not removable.

On the other hand, if we take eqn (3.26) to be describing the patch $0 < r < 2m$, then it is seen that as r tends to zero, the curvature scalar

$$R^{ijkl}R_{ijkl} = \frac{48m^2}{r^6},$$

diverges and it follows that the point $r = 0$ is a real space-time singularity. It is not possible to extend the space-time across this singularity in a continuous manner.

Such a maximal extension of the manifold (3.26) with $2m < r < \infty$ was obtained by Kruskal (1960) and Szekeres (1960). We describe this procedure below, which uses suitably defined advanced and retarded null coordinates. Using the condition for null geodesics, that is,

$$g_{ij}K^iK^j = 0,$$

the radial null geodesics in the Schwarzschild space-time (3.26) are given by

$$\left(\frac{dt}{dr}\right)^2 = \left(\frac{r}{r-2m}\right)^2. \tag{3.27}$$

Define r^* as

$$r^* \equiv \int \frac{dr}{1-2m/r} = r + 2m\log(r/2m - 1). \tag{3.28}$$

The radial null geodesics above satisfy

$$t = \pm r^* + \text{const.} \tag{3.29}$$

The null coordinates u and v are now defined by

$$u = t - r^*, \qquad v = t + r^*. \tag{3.30}$$

Thus, $r^* = (v-u)/2$, which is given in terms of r by eqn (3.28). Now r can be viewed as defined implicitly in terms of u and v. Then, using above equations, the metric (3.26) can be written as

$$ds^2 = -\frac{2me^{-r/2m}}{r}e^{(v-u)/4m}dudv + r^2d\Omega^2. \tag{3.31}$$

Now $r \to 2m$ corresponds to $u \to \infty$ or $v \to -\infty$. Define new coordinates U and V now by

$$U = -e^{-u/4m}, \qquad V = e^{v/4m}, \tag{3.32}$$

which gives the non-angular part of the metric as

$$ds^2 = -\frac{32m^3e^{-r/2m}}{r}dUdV. \tag{3.33}$$

There is no singularity now at $U = 0$ or $V = 0$, which corresponds to the value $r = 2m$. Now a final transformation of the form $T = (U+V)/2$ and $X = (V-U)/2$ gives the Schwarzschild metric in the Kruskal–Szekeres form

$$ds^2 = \frac{32m^3e^{-r/2m}}{r}(-dT^2 + dX^2) + r^2d\Omega^2. \tag{3.34}$$

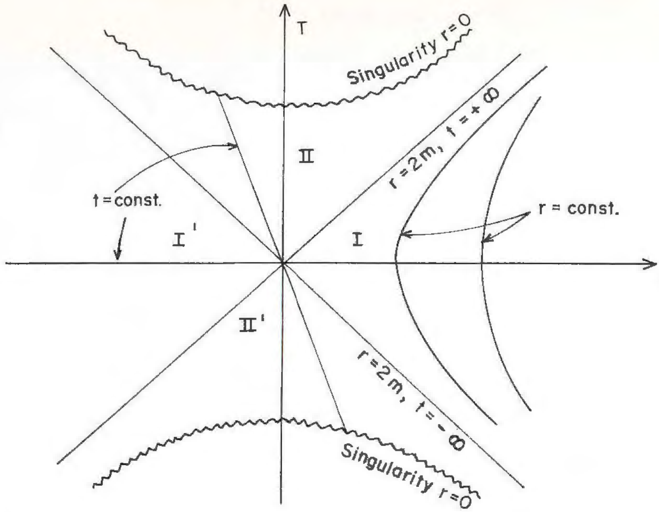

Fig. 13 The Kruskal extension of the Schwarzschild geometry.

The coordinate transformation between the original coordinates (t, r) and new coordinates (T, X) is given by

$$X^2 - T^2 = \left(\frac{r}{2m} - 1\right) e^{r/2m}, \tag{3.35}$$

$$t = 4m \tanh^{-1}(T/X). \tag{3.36}$$

The quantity r in eqn (3.34) is determined implicitly by eqn (3.35). The condition $r > 0$ specifies the allowed range of coordinates, which is $X^2 - T^2 > -1$.

The structure of this maximal Schwarzschild manifold is shown in Fig. 13, which is also called the Kruskal extension of the Schwarzschild space-time. The radial null geodesics are 45° lines in the X, T coordinates.

The physical singularity at $r = 0$ corresponds to the values $X = \pm(T^2 - 1)^{1/2}$, and we note that there is no singularity in the metric now at $r = 2m$. The original Schwarzschild solution for $r > 2m$ corresponds to the region I here which is interpreted as the exterior gravitational field of a collapsing body. Region I is asymptotically flat and so is region I', which has identical properties as region I. (Note however, that the Kruskal representation is not best suited to study asymptotic properties and it is best to use a conformal compactification of the metric (3.26) for that purpose.) There is no causal communication between regions I and I'; any observer or photon from region I either goes away to infinity or crosses the null line $X = T$ and enters region II. Once a radially infalling observer

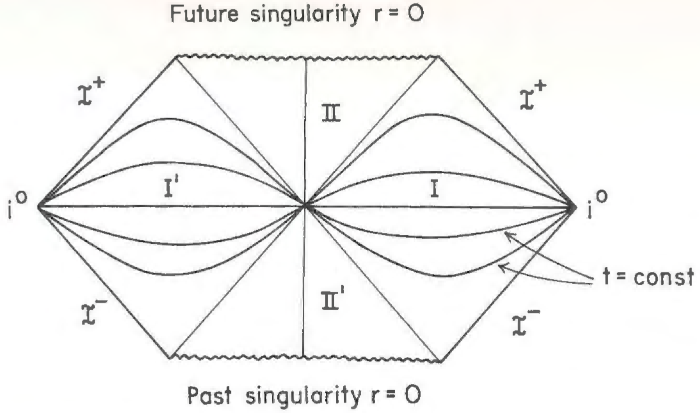

Fig. 14 A conformal picture of the Kruskal geometry.

is inside region II, there is no escape from it and within a finite proper time the observer must fall into the singularity at $X = (T^2 - 1)^{1/2}$. Any light signal emitted in region II also must fall into the singularity and it can never cross into region I. Hence, region II is termed a *black hole*. The region II' has time reversed properties of region II and is also called a *white hole*. A particle emitted by the singularity at $X = -(T^2 - 1)^{1/2}$ must leave this region within a finite proper time. Each point of Fig. 13 represents a two-sphere in the space- time. If a source at a point p in the region $r > 2m$ emits a flash of light, there will be two two-spheres formed, one by the outgoing wave front and the other by ingoing wave front. The outgoing sphere will have a greater area as compared to the ingoing one. However, if the source p lies in the limit $r < 2m$, both the outgoing and ingoing spheres will have areas less than that of p. Then we say that p is a *closed trapped surface*. Such surfaces play an important role towards showing the existence of space-time singularities discussed in Chapter 5. Just as in the Minkowski case, one could construct a conformal compactification of the above Kruskal extension of the Schwarzschild geometry, which is more convenient as far as the investigation of asymptotic structure is concerned. Such a conformal diagram of Kruskal geometry is given in Fig. 14.

Whereas the regions I and II of the extended Schwarzschild manifold have a clear physical interpretation as discussed above, the physical relevance of regions I' and II' is not very obvious. Neither is it possible to rule them out easily. While points in region II' are time-reversed closed trapped surfaces, region I' is another asymptotically flat universe on the other side of the Schwarzschild 'throat'. This is clear from considering the spatial geometry of the hypersurface $t = 0$. The two-spheres $r =$ const. are almost flat Euclidian for large values of r, but for small r, their area

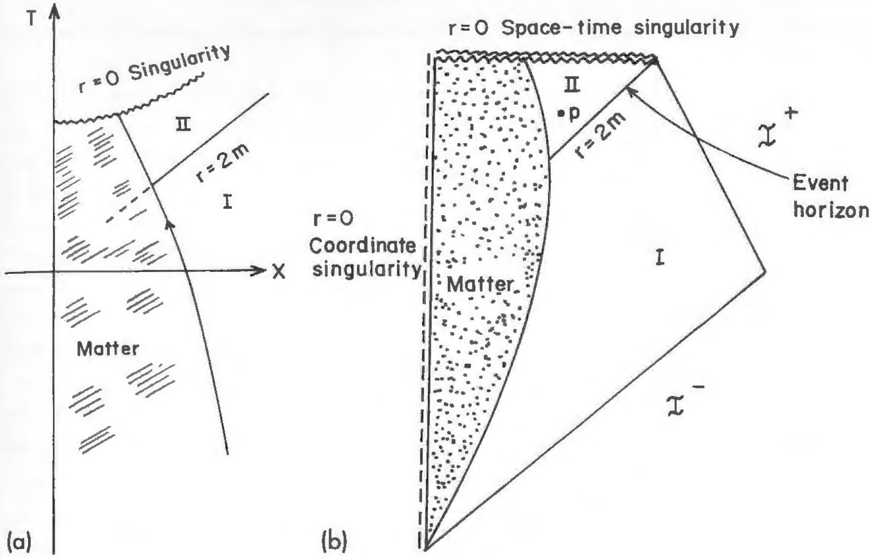

Fig. 15 (a) Complete gravitational collapse of a homogeneous dust cloud represented in the Kruskal picture. The regions I' and II' of Fig. 13 are completely covered by the matter now and so is a part of the region II. The uncovered portion of region II represents the formation of a black hole. The points in region I and II of the figure represent two-spheres in the space-time and any point, such as p in Fig. 15b, within the horizon is a closed trapped surface. (b) A Penrose diagram for this collapse scenario.

decreases to a minimum corresponding to that of the value $r = 2m$, and then it increases again as the two spheres expand in the other region of asymptotically flat three-space. However, if we consider a complete gravitational collapse of a spherically symmetric homogeneous dust cloud, the regions I' and II' are no longer relevant as they are replaced by the interior metric which is not vacuum Schwarzschild, T_{ij} being non-zero there. The situation is shown in Fig. 15a, and a conformal diagram of such a collapse is given in Fig. 15b.

The uncovered portions of regions I and II represent the vacuum Schwarzschild geometry exterior to the collapsing matter. The portion of region II indicates that a Schwarzschild black hole is always produced in the complete gravitational collapse which fully covers the resulting space-time singularity of infinite curvature and density. This situation has a great significance for the cosmic censorship hypothesis and the black hole formation to be discussed in Chapter 6. The interior metric in this case is exactly that of a closed Friedmann model which we will consider in Section 3.6.

In the extended Schwarzschild manifold, the surface $r = 2m$ is a null hypersurface and each point there is a two-sphere of area $16\pi m^2$. Note that in eqn (3.26), the component $g_{00} = (1 - 2m/r) > 0$ for $r > 2m$, however, $g_{00} < 0$ for $r < 2m$. Thus, it is no longer possible to use t as a time coordinate as the coordinates t and r reverse their roles and space-time is no longer static. Thus, $r = 2m$ surface is called a 'static limit' as well. The vector $\partial/\partial t$ with components $\xi^i = \delta^i_0 = (1, 0, 0, 0)$ gives the time translation, leaving the g_{ij} unchanged as it does not involve the time coordinate. Thus, ξ is a Killing vector which leaves the space-time geometry unchanged. We have $\xi^2 = g_{ij}\xi^i\xi^j = g_{00}$ and for the Schwarzschild metric, ξ^2 vanishes on $r = 2m$. Hence, at the static limit the timelike Killing vector becomes null. In the Kruskal diagram also it is seen that ξ^i vanishes at $X = T = 0$ and this leads to the odd labeling of lines $X = \pm T$ as '$t = \pm\infty$'.

The Schwarzschild geometry provides an illustration of the basic principle which Einstein used to formulate his gravitation theory, namely that matter tells the space-time in its vicinity how to curve. To see this, consider the Schwarzschild solution in a spacelike surface $t =$ const. and in the equatorial plane $\theta = \pi/2$. The metric of this two-dimensional curved surface is described by the metric

$$ds^2 = \frac{dr^2}{(1 - 2m/r)} + r^2 d\phi^2.$$

The geometry of such a curved surface can be visualized as embedded in the ordinary Euclidian space. Here, the region $0 < r < r_b$ is to be considered as filled by the matter which represents the spherical star with a boundary at $r = r_b$, and the curved surface would then represent the geometry outside such a star.

Consider now a static observer along a Killing direction, for whom the four-velocities are $u^i = \xi^i / \mid \xi \mid$. Suppose now a static source with four-velocity u^i_1 emits a photon with four-momentum p^i (so $p_{i;j}p^j = 0$ with a suitable parametrization) and is observed by a static observer with four-velocity u^i_2 (Fig. 16). Now, take the directional derivative of $\xi_i p^i$ along the geodesic tangent p^i,

$$(\xi_i p^i)_{;j} p^j = \xi_{i;j} p^i p^j + \xi_i p^i_{;j} p^j = 0. \tag{3.37}$$

The first term vanishes because ξ^i is a Killing vector and the second term vanishes because of the geodesic equation. The ratio of energies measured at these two points by static observers is given by

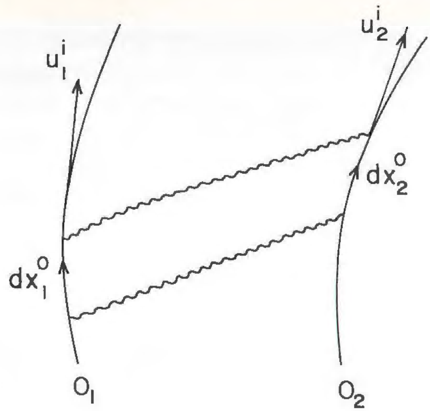

Fig. 16 The static source O_1 emits light rays which are received by the static observer O_2.

$$\frac{E_1}{E_2} = \frac{(u^i p_i)_1^{1/2}}{(u^i p_i)_2^{1/2}}. \tag{3.38}$$

Using $u^i = \xi^i / \mid \xi \mid$ and the implication of eqn (3.37) that $\xi^i p_i = \text{const.}$ along the geodesic, we get

$$\frac{E_1}{E_2} = \frac{(\xi^i \xi_i)_1{}^{1/2}}{(\xi^i \xi_i)_2{}^{1/2}}. \tag{3.39}$$

Since $\xi^2 = g_{00}$, this is the gravitational red-shift formula for a static source and observer in terms of the metric components. It is now seen that if the observer remains at a finite radius but the source approaches $r = 2m$, the red-shift tends to infinity. Thus, as a particle falls into the black hole approaching $r = 2m$, the light rays emitted by it are infinitely red-shifted as observed by a distant static observer in the outside space-time.

As pointed out above, the Schwarzschild space-time is asymptotically flat. For a source situated outside $r = 2m$, part of the photon trajectories emitted with decreasing r values will enter the black hole and fall into the singularity. All other null geodesics will escape to infinity to intersect $\mathcal{I}^+$. If a source is located below $r = 2m$, no null geodesic can come out of the black hole and all end up in the singularity in future. As in the case of Minkowski space-time, we now work out the light cone cuts of future null infinity from an arbitrary apex in the Schwarzschild region $r > 2m$ (Joshi, Kozameh and Newman, 1983). This process leads to obtaining all the null geodesics and the full light cone from a given point in the space-time in Section 4.7. Such null trajectories in Schwarzschild geometry are of considerable importance as they are used to verify the general relativity theory experimentally by means of effects such as bending of light rays near a star, the time delay of light and other such effects.

The Schwarzschild metric in (u, r, θ, ϕ) coordinates, where $u = t - r - 2m\log(r - 2m)$ is the retarded time, is given as

$$ds^2 = -\left(1 - \frac{2m}{r}\right) du^2 - 2dudr + r^2(d\theta^2 + \sin^2\theta\, d\phi^2). \tag{3.40}$$

As in the case of the Minkowski space-time, we make necessary coordinate transformations, use stereographic coordinates $\zeta, \bar{\zeta}$ and conformally transform the metric by $\Omega = r^{-1} = \sqrt{\ell}$ which gives

$$d\bar{s}^2 = \Omega^2 ds^2 = -4(\ell^2 - 2\sqrt{2}m\ell^3)du^2 + 4dud\ell + \frac{d\zeta d\bar{\zeta}}{P_0^2}. \tag{3.41}$$

The new coordinate ℓ is now finite at infinity and $\mathcal{I}^+$ is described by the hypersurface $\ell = 0$, which corresponds to $r = \infty$. The Lagrangian for the geodesics is written as

$$L = 2(\ell^2 - 2\sqrt{2}m\ell^3)\dot{u}^2 - 2\dot{u}\dot{\ell} - \frac{\dot{\zeta}\dot{\bar{\zeta}}}{2P_0^2}, \tag{3.42}$$

where the dot denotes differentiation with respect to an affine parameter s along null geodesics. The equations for null geodesics are then given as:

$$\begin{aligned}
2(\ell^2 - 2\sqrt{2}m\ell^3)\dot{u} - \dot{\ell} &= 1,\\
\ddot{u} + 2(\ell - 3\sqrt{2}m\ell^2)\dot{u} &= 0,\\
\ddot{\zeta}(1 + \zeta\bar{\zeta}) - 2\bar{\zeta}\dot{\zeta}^2 &= 0,\\
\ddot{\bar{\zeta}}(1 + \zeta\bar{\zeta}) - 2\dot{\bar{\zeta}}^2\zeta &= 0,\\
4(\ell^2 - 2\sqrt{2}m\ell^3)\dot{u}^2 - 4\dot{u}\dot{\ell} - \frac{\dot{\zeta}\dot{\bar{\zeta}}}{P_0^2} &= 0,
\end{aligned} \tag{3.43}$$

where the last equation corresponds to $ds^2 = 0$. Though, in principle, all the null geodesics of the space-time are obtained from eqns (3.43), we first consider only those in the equatorial plane $\theta = \pi/2$. From a fixed apex this yields an S^1 worth of null geodesics. Using these and the spherical symmetry of the space-time we can generate all the null geodesics from an arbitrary apex by a rigid rotation. For $\theta = \pi/2$ we have $\zeta = e^{i\phi}$, which gives

$\ddot{\phi} = 0$, that is $\dot{\phi} = b$. Further, the equation for $\ddot{u}$ follows as an identity from other equations. Combining the first and the last of the above equations we can then write

$$\begin{aligned} \dot{u} &= \frac{1+\dot{\ell}}{2(\ell^2 - 2\sqrt{2}m\ell^3)}, \\ ds &= \pm\frac{d\ell}{\sqrt{A}}, \\ d\phi &= \pm\frac{bd\ell}{\sqrt{A}}, \end{aligned} \tag{3.44}$$

where the cubic A is given by

$$A = 2\sqrt{2}mb^2\ell^3 - b^2\ell^2 + 1.$$

Before integrating the above, we note that the null rays coming from an arbitrary apex are divided into two sets (the two sheets of A) , that is, those given initially by $\dot{\ell} < 0$ and $\dot{\ell} > 0$. For the first set, the geodesics continue with a decreasing ℓ (increasing r) until intersection with $\mathcal{I}^+$. For the rays which begin with $\dot{\ell} > 0$ that is, those rays with initially increasing ℓ (decreasing r), some reach a maximum ℓ (when $A = 0$) and then begin to move outwards, and eventually also intersect $\mathcal{I}^+$. For others, depending on the value of the impact parameter b, they continue towards increasing ℓ and eventually fall within the horizon and do not reach $\mathcal{I}^+$. We shall not be concerned with the later rays.

For a fixed apex (say at $\ell = \ell_0 < 1/3\sqrt{2}m$), the null rays, on each sheet, are characterized by the value of the impact parameter b. For the first sheet, the range of b is from $b = 0$ to a maximum b_m, where the b_m is determined by $A = b_m^2(2\sqrt{2}m\ell_0^3 - \ell_0^2) + 1 = 0$, which yields the value for b_m. For the second sheet ($\dot{\ell} > 0$), the range is again from some b_m to $b = 0$, but now there is a critical value b_c such that for all $b < b_c$ the rays continue past the horizon. To determine b_c, we want the smallest b so that A has a real positive root ℓ_c. By plotting A against ℓ it is easily calculated that ℓ_c is a double root and $\ell_c = 1/3\sqrt{2}m$ with $b_c = 3\sqrt{6}m$. Thus, on the second sheet, the range for b is $b_c < b < b_m$. Note that a ray beginning at $\ell = \ell_0$ with $b = b_c$ approaches asymptotically the well-known (unstable) orbit $\ell = \ell_c$.

We can now integrate eqns (3.44) from a fixed apex to the value $\ell = 0$, which represents the future null infinity, and this gives part of the light cone cut at the infinity. Using that and the spherical symmetry of the space-time gives the full light cone cut of the future null infinity.

3.3 The Kerr solution

As pointed out in the previous section, the final configuration resulting from a complete spherically symmetric gravitational collapse is described by the Schwarzschild geometry in the exterior of the collapsing dust cloud configuration. Even when the collapse is non-spherical and the outside geometry is time-dependent, one would expect that ultimately the geometry will settle to a stationary final state. Further, all astronomical bodies rotate and so one would not expect the exterior solution to be exactly spherically symmetric. The Kerr space-time incorporates these features and models the exterior geometry outside a rotating object and also a rotating black hole (Kerr, 1963). The metric is written in the Boyer–Lindquist (1967) coordinates (t, r, θ, ϕ) as

$$\begin{aligned} ds^2 = &-\left(1 - \frac{2mr}{\Sigma}\right) dt^2 - \frac{4amr\sin^2\theta}{\Sigma} dtd\phi \\ &+ \frac{\Sigma}{\triangle} dr^2 + \Sigma d\theta^2 + \left(r^2 + a^2 + \frac{2mra^2 sin^2\theta}{\Sigma}\right) \sin^2\theta \, d\phi^2, \end{aligned} \tag{3.45}$$

where

$$\Sigma \equiv r^2 + a^2\cos^2\theta, \qquad \triangle \equiv r^2 - 2mr + a^2. \tag{3.46}$$

The parameter m is interpreted as earlier as the gravitational mass of the system and if J represents the total angular momentum of space-time then $a = J/m$, that is, it is the angular momentum per unit mass.

The metric components here are again time-independent and the space-time admits a timelike Killing vector $\xi^i = \delta^i_0$. The g_{ij} components are also independent of the coordinate ϕ. Hence the space-time is axially symmetric and admits an axial Killing vector $\psi^i = \delta^i_3$. The Kerr geometry is *stationary* as it admits a timelike Killing vector field. The stationarity and the axisymmetry property of the solution are related to the fact that the solution admits both $\partial/\partial t$ and $\partial/\partial\phi$ as Killing vector fields. In such a case, the solution is invariant under rotation around the z-axis, that is, $\theta = 0$ and the orbits of the Killing vector $\partial/\partial\phi$ are circles. Also, the solution is invariant under the simultaneous inversion of the t and ϕ coordinates given by $t \to -t$ and $\phi \to -\phi$ implying that the geometry might arise due to a spinning source. There are cross terms in the metric such as $dtd\phi$ which cannot be transformed away and this feature incorporates the rotation inherent in the space-time. In a stationary space-time the metric components are time independent but the rotational terms such as the above are included. Again, as $r \to \infty$, the metric components tend to the Minkowskian values, which indicates that the space-time is asymptotically flat. When we set $a = 0$, not only the rotational term vanishes but the metric (3.45) reduces to the Schwarzschild space-time.

In order to investigate the singularity structure of the Kerr space-time, we can evaluate the curvature invariant $R^{ijkl}R_{ijkl}$ and it can be shown that the only non-removable curvature singularity lies at

$$\Sigma = r^2 + a^2 \cos^2 \theta = 0.$$

This would happen when $r = \cos\theta = 0$ and the singularity has the structure of a ring of radius a which lies in the equatorial plane $z = 0$.

The square of the timelike Killing vector, $\xi^i = \delta^i_0$, is given by

$$\xi^2 = \xi^i \xi_i = g_{00} = (r^2 - 2mr + a^2 \cos^2 \theta)\Sigma^{-1}. \tag{3.47}$$

Analogous to the static limit surface in the Schwarzschild geometry, the *stationary limit surface* in Kerr geometry is given as a surface where the Killing vector ξ^i becomes null. This is interpreted as the surface where the coordinate t changes its nature from being timelike to become spacelike. The stationary limit is thus given by

$$\xi^2 = 0 = r^2 - 2mr + a^2 \cos^2 \theta,$$

which corresponds to the values

$$r = m \pm (m^2 - a^2 \cos^2 \theta)^{1/2}. \tag{3.48}$$

Of the two values given above, generally the outer stationary limit (given by choosing the positive sign) is relevant for most of the discussions. In the Kerr space-time we consider the stationary sources and observers rather than the static ones. A stationary observer could be defined up to the stationary limit, who follows the Killing direction with four-velocities $u^i = \xi^i / \mid \xi \mid$. A discussion similar to the Schwarzschild case now shows that as the source particle approaches the stationary limit, the phenomena of infinite red-shift occurs.

We note that when $a = 0$, the outer stationary limit surface coincides with the static limit surface $r = 2m$ of the Schwarzschild space-time, which is null. However, the stationary limit surface in the Kerr geometry is not a null surface. It is a timelike surface except at two points on the axis where it is null. Hence, it can be crossed by the particles in either ingoing or outgoing directions and it is not the one way membrane that defines the Kerr black hole. The significance of this surface is that exterior to it timelike particles can travel on the orbit of the timelike Killing vector ξ^i and so could remain at rest with respect to infinity.

To locate the black hole surface, we note that the metric has a singularity, similar to the $r = 2m$ Schwarzschild case, when $\triangle = 0$. The two values of r given by this equation are

$$r_\pm = m \pm (m^2 - a^2)^{1/2}. \tag{3.49}$$

Assuming that $a^2 < m^2$, this gives rise to two null event horizons. We will be concerned with the outer value $r = r_+$. One could view eqn (3.49) as the equation of a surface in the space-time. Working out the surface normal and using the fact that $\Delta = 0$ it follows that $r = r_+$ is a null surface. Thus, a particle which crosses it in future direction cannot return again to the same region. It forms the boundary of the region in the space-time from which particles can escape to the future null infinity. Thus the surface $r = r_+$ is the *event horizon* or the *black hole surface* for the future null infinity. It follows that just as the Schwarzschild solution is regular in two different disconnected regions in the Schwarzschild coordinates, the Kerr solution is regular in three regions given by, $0 < r < r_-$, $r_- < r < r_+$, and $r_+ < r < \infty$. This suggests the possibility of the maximal analytic extension just as in the Schwarzschild case.

Of course, such a black hole forms only when $0 < a^2 < m^2$. For $a > m$, there is no black hole and the genuine curvature singularity at $r = 0$ is visible to external observers. Such a singularity is called a *naked singularity*, which is not covered by the event horizon. A major unsolved problem in general relativity to day is Penrose's *cosmic censorship hypothesis*, which states that a naked singularity such as above will never result in a complete gravitational collapse from well-behaved initial conditions. In the case of rotating stars, we are certainly not aware of any mechanism which will speed up a star with $a < m$ to a situation when $a > m$. In the present case, it can be seen by transforming to the Kerr–Schild coordinates that the singularity at $r = 0$ is not a point but actually a ring (see for example, Hawking and Ellis (1973), for details). Further, causality violation is seen to take place in the vicinity of this naked singularity. Near the singularity, the vector $\partial/\partial\phi$ is timelike and hence the circles with $t, r, \theta =$ const. are closed timelike curves which can be made to pass through any point of the space-time (Carter, 1968). Thus the causality violation is global.

The maximal analytic extension of the Kerr space-time for the case $a > m$ is given by Carter (1968). In the case of $a^2 < m^2$, the maximal analytic extension can be obtained by extending the metric (3.45) across the surfaces $r = r_+$ and $r = r_-$. This was obtained by Carter (1968) where the situation is more complicated than the case when $a^2 > m^2$. A notable difference is now there is no longer a global causality violation. However, in the region $-\infty < r < r_-$ containing the ring singularity, there are closed timelike curves through all the points. No causality violation occurs in the region given by $r_- < r < \infty$. We note that for the case $a = m$, the values r_+ and r_- coincide. The outer stationary limit r_0 always lies outside of the event horizon r_+ except at two points on the axis where it is null and coincides with the surface $r = r_+$.

The region bounded by r_+ and r_0 is called the *ergosphere* in which

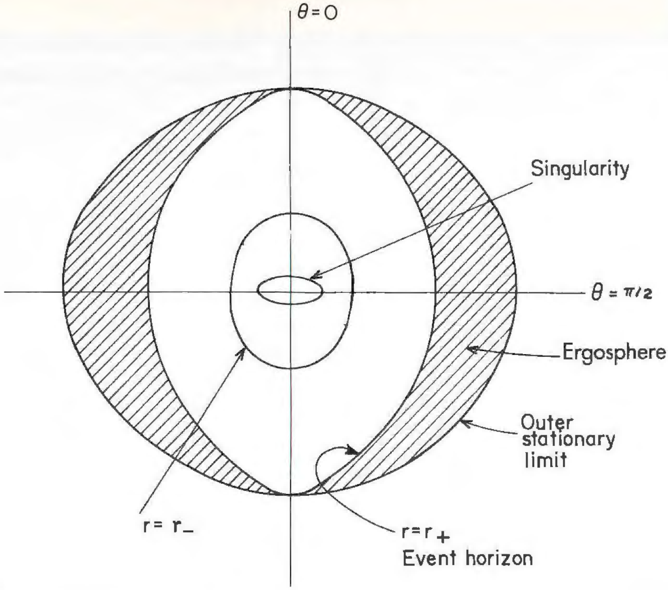

Fig. 17 The stationary limit surface and event horizon in Kerr solution with $0 < a^2 < m^2$.

the Killing vector ξ^i is spacelike. Thus, in this region, no timelike particle or photon can travel along an orbit of the Killing vector and remain at rest with respect to infinity. In fact, it is possible for particles from the ergosphere to cross the stationary limit and reach to infinity. However, from the surface r_+ or from inside it is not possible to escape to infinity for any timelike or null curve. These features are shown in Fig. 17.

The discussion of non-spacelike geodesics in Kerr geometry is somewhat complicated in the absence of spherical symmetry. However, the existence of the timelike Killing vector ξ^i and the axial Killing vector ψ^i yields a conserved energy E and conserved angular momentum L, which could be evaluated from the derivatives of the Lagrangian with respect to $\dot{t}$ and $\dot{\phi}$ respectively and are given as

$$E = \left(1 - \frac{2mr}{\Sigma}\right)\dot{t} + \frac{2mar\sin^2\theta}{\Sigma}\dot{\phi}, \tag{3.50}$$

$$L = -\frac{2mar\sin^2\theta}{\Sigma}\dot{t} + \frac{(r^2+a^2)^2 - \Delta a^2\sin^2\theta}{\Sigma}\sin^2\theta\dot{\phi}, \tag{3.51}$$

where a dot denotes a derivative with respect to the affine parameter. Further, we have

$$g_{ij}\dot{x}^i\dot{x}^j = -A, \tag{3.52}$$

where $A = 1$ for timelike geodesics and $A = 0$ for null geodesics. Eliminating $\dot{t}$ and $\dot{\phi}$ from the above equations and confining to null geodesics in the equatorial plane $\theta = \pi/2$ gives

$$\frac{\dot{r}^2}{2} + F(E, L, r) = 0, \tag{3.53}$$

where the function F is given by

$$F = \frac{L^2}{2r^2} - \frac{E^2}{2}\left(1 + \frac{a^2}{r^2}\right) - \frac{m}{r^3}(L - aE)^2. \tag{3.54}$$

The above could then be analysed further to get the properties of photon trajectories in the equatorial plane.

We shall be particularly interested in general non-equatorial photon trajectories in Kerr space-times in Section 4.7. Such null geodesics for the Schwarzschild space-time were discussed in the previous section. The information provided by eqns (3.50) and (3.51) is not enough for this purpose. However, in this case it is possible to use the separability of the Hamilton-Jacobi equation for geodesics, as was shown by Carter (1968), to integrate the geodesic equations in an explicit manner. In fact, as pointed out by Walker and Penrose (1970), the Kerr metric admits a Killing tensor K_{ij} and hence an additional constant of motion given by $K_{ij}\, u^i u^j$, which can be used to integrate the geodesic equations. For the derivation of general equations of non-spacelike geodesics in Kerr space-time and for a complete discussion on topics such as the separability of the Hamilton–Jacobi equation, the Penrose process for energy extraction from the ergosphere and electromagnetic and gravitational perturbations of a Kerr black hole, we refer to Chandrasekhar (1983).

3.4 Charged Schwarzschild and Kerr geometries

The final state of a complete gravitational collapse, either spherically symmetric or non-spherically symmetric, is expected to be a vacuum space-time which incorporates the rotation and also possibly the electromagnetic fields associated with the object. It is possible that the charge associated with an astrophysical object could be quickly neutralized by the surrounding plasma. However, in any case it will be of interest to obtain all solutions of Einstein–Maxwell equations which could describe stationary collapsed configurations with charge.

The charged generalization of Schwarzschild geometry, representing the space-time outside a spherically symmetric electrically charged body, is called the Reissner–Nordström metric. This is a static and asymptotically flat solution of the Einstein–Maxwell equations $G_{ij} = 8\pi T_{ij}$. Here T_{ij} is the Maxwell energy-momentum tensor which is trace-free and the Ricci tensor vanishes. Further, in source-free regions, the Maxwell tensor F_{ij} satisfies the Maxwell equations

$$\nabla_j F^{ij} = 0, \quad \partial_{[i} F_{jk]} = 0.$$

In (t, r, θ, ϕ) coordinates the metric is written as

$$ds^2 = -\left(1 - \frac{2m}{r} + \frac{e^2}{r^2}\right) dt^2 + \left(1 - \frac{2m}{r} + \frac{e^2}{r^2}\right)^{-1} dr^2 + r^2 d\Omega^2, \quad (3.55)$$

where $d\Omega^2 = d\theta^2 + \sin^2\theta\, d\phi^2$, m is the gravitational mass and e the electric charge of the body. The stress-energy tensor T_{ij} no longer vanishes in the exterior represented by this solution but it corresponds to the electromagnetic field in the space-time resulting from the charge of the body. In fact, we can again establish here an analogy to Birkhoff's theorem, namely that a spherically symmetric solution of the Einstein–Maxwell equations is necessarily static.

It is seen by taking the limit as $r \to \infty$ that the solution is asymptotically flat and if we regard it as a solution for all r, it is clear that there is a genuine curvature singularity at $r = 0$ which cannot be removed by any coordinate transformations, and which could be regarded as the point charge which produces the field. When $e^2 > m^2$, the metric is non-singular everywhere except at the $r = 0$ singularity. If $e^2 \leq m^2$, the metric has further coordinate singularities given by

$$r_\pm = m \pm (m^2 - e^2)^{1/2}. \quad (3.56)$$

However, the metric is regular in the regions $0 < r < r_-$, $r_- < r < r_+$ and $r_+ < r < \infty$. When $m^2 = e^2$, we have $r_- = r_+$ and when $e = 0$, the solution reduces to the Schwarzschild solution.

The maximal analytic extension of the metric (3.55) was obtained by Graves and Brill (1960). Similar to the Schwarzschild case, one could write eqn (3.55) in double null coordinates to remove the singularities at $r_\pm$ by means of a further coordinate transformation.

The Penrose diagram for this maximally extended metric is shown in Fig. 18 for the case $e^2 < m^2$. Similar diagrams can be given for $e^2 \geq m^2$ situation (see for example, Hawking and Ellis, 1973). A notable feature in each case is the appearance of a timelike singularity at $r = 0$, unlike the Schwarzschild case where the singularity is spacelike. This singularity

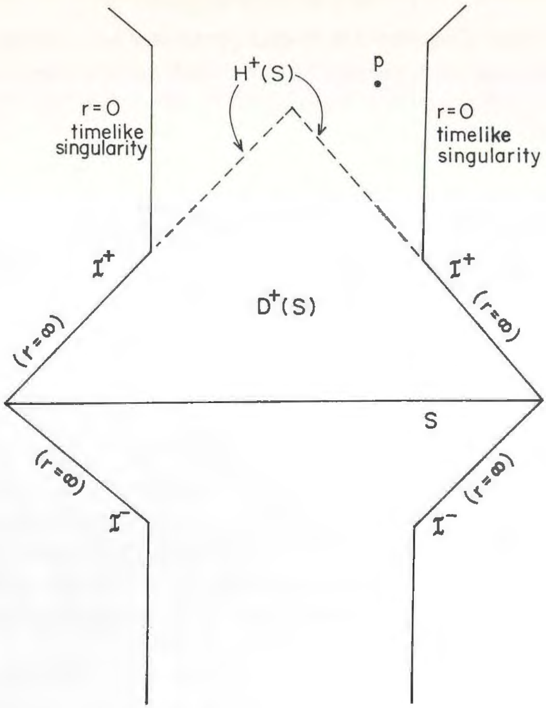

Fig. 18 A timelike singularity forms towards the future of a partial Cauchy surface S as shown in the Penrose diagram of the Reissner–Nordström space-time ($e^2 < m^2$). There are past-directed non-spacelike curves from events such as p which do not meet S but terminate at the naked singularity in the past.

is repulsive in the sense that no timelike geodesics meet there but non-geodetic timelike curves and radial null geodesics can.

The timelike singularity at $r = 0$ is a naked singularity in the sense that there are future directed non-spacelike curves in the space-time which have their past end points at the singularity. Could this be regarded as a violation of cosmic censorship hypothesis stated in Section 3.3? The point is, the Reissner–Nordström metric, taken as a solution for all values of r, admits the naked singularity at $r = 0$ for all times, that is, it is persistently naked singular at all epochs. What one desires to rule out by means of a censorship hypothesis is that naked singularities should not develop from non-singular initial data. Thus, one would like to consider a space-time which admits a non-singular spacelike hypersurface with regular initial conditions. Consider, for example, the evolution from such a surface because of the gravitational collapse of matter with a reasonable equation of state. If a naked singularity results this way in the future of a non-singular initial surface, this may be considered as a counter-example to the cosmic censorship hypothesis. Now the Reissner–Nordström solution or the Kerr solution with $a^2 > m^2$ discussed above cannot be generated as the end product of gravitational collapse by any known mechanism and so they may not be regarded as counter-examples of the cosmic censorship hypothesis.

Another point to note here is, unlike the Schwarzschild space-time, the Reissner–Nordström case is not globally hyperbolic in the sense that it will not admit a Cauchy hypersurface. This is because, given any spacelike hypersurface, there are timelike or null curves which run into the singularity and do not meet this surface. This is a violation of cosmic censorship principle in its strong form which says that the space-time must be globally hyperbolic. Such a space-time admits Cauchy horizons which mark the boundary of the region predictable from a partial Cauchy surface indicating the break down of global hyperbolicity.

An important question to be asked here is are these Cauchy horizons stable? If they are not, then again this may not be regarded as a serious counter-example to cosmic censorship. Stability of such Cauchy horizons have been analysed by several authors (for example, Simpson and Penrose, 1973; Chandrasekhar and Hartle, 1983) in various space-times including the Reissner–Nordström case. It is argued that a small perturbation in the initial data which extends to infinity on a partial Cauchy surface without violating asymptotic flatness should give rise to an 'infinite blueshift' singularity on the Cauchy horizon. Such a behaviour is shown to occur for the case of Reissner–Nordström Cauchy horizon by Chandrasekhar and Hartle (1982). Thus, even though violation of global hyperbolicity occurs in certain situations, it is not clear how stable and generic this behaviour is.

It is possible to join two Reissner–Nordström solutions with different masses m_1 and m_2 and charges e_1 and e_2 across charged spherical shells of massless matter (Dray and Joshi, 1990). This procedure also allows one to join a Reissner–Nordström space-time to a Schwarzschild space-time or with a flat space solution, and conservation laws can be used to define the charge and energy of the shells.

We now consider briefly the charged generalization of Kerr solution which is called the Kerr–Newman metric. This was obtained by Newman et al. (1965) and is a three-parameter family characterized by the parameters $m, a,$ and e which are called the mass, angular momentum, and charge respectively. The metric is obtained by simply redefining the quantity $\triangle$ in the Kerr solution (3.45) as

$$\triangle \equiv r^2 - 2mr + a^2 + e^2. \tag{3.57}$$

The global properties of this family are very similar to the uncharged Kerr version. When $e = 0$, the electromagnetic vector potential vanishes, $T_{ij} = 0$ and the metric reduces to the vacuum Kerr family. When $a = 0$, we recover the Reissner–Nordström solution and for $a = e = 0$ the metric reduces to the Schwarzschild case. Thus, this three-parameter family covers all known stationary black hole solutions. The space-time is stationary and axisymmetric with the same Killing vectors ξ^i and ψ^i as the Kerr geometry and is also asymptotically flat. We will be interested later (Section 4.8) in the asymptotic behaviour of null geodesics and the light cone from an arbitrary apex in this space-time. As stated in the previous section, the first integrals of the null geodesic equations can be obtained using the Hamilton–Jacobi method (Carter, 1968), and using null coordinates (u, r, θ, ϕ), they are given by

$$\begin{aligned}
\rho^2\dot{\theta} &= \sqrt{\Theta},\\
\rho^2\dot{r} &= \sqrt{R},\\
\rho^2\dot{u} &= (aL - a^2\sin^2\theta) + \frac{(r^2+a^2)(\sqrt{R}+P)}{\Delta},\\
\rho^2\dot{\phi} &= \left(\frac{L}{\sin^2\theta} - a\right) + \frac{a(\sqrt{R}+P)}{\Delta},
\end{aligned} \tag{3.58}$$

where

$$\begin{aligned}
&\rho^2 = r^2 + a^2\cos^2\theta, \quad \Delta = r^2 - 2mr + a^2 + e^2,\\
&\Theta = Q - \cos^2\theta(L^2/\sin^2\theta - a^2); \quad Q = K - (L-a)^2,\\
&P = (r^2+a^2) - La,\\
&R = P^2 - \Delta K.
\end{aligned} \tag{3.59}$$

Here, a dot denotes the derivative with respect to an affine parameter along the null geodesics and L and K are constants of integration.

3.5 The Vaidya radiating metric

The Vaidya metric, which is also called the radiating Schwarzschild metric (Vaidya, 1943, 1951, 1953) describes the geometry outside a spherically symmetric star when the emission of radiation from the star is included. Just as astrophysical bodies are found to be rotating, they also radiate energy in the form of electromagnetic radiation. The Schwarzschild solution does not describe this situation as it corresponds to an empty exterior given by $T_{ij} = 0$; neither does the Reissner–Nordström case which corresponds

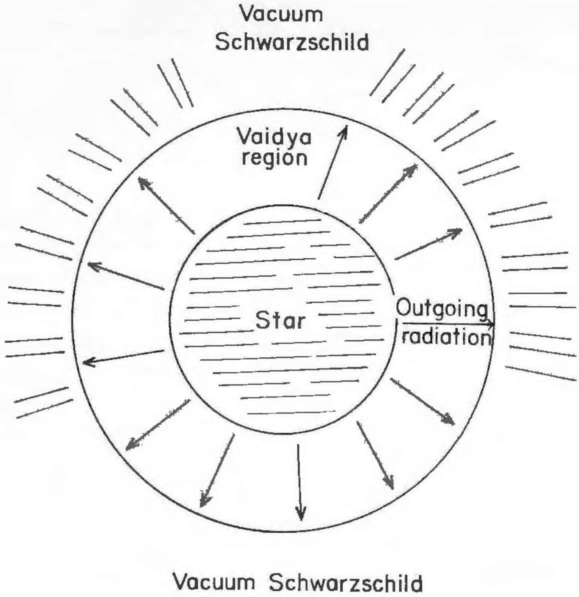

Fig. 19 A schematic diagram for the radiating star configuration. The interior of the star is matched with the Vaidya metric outside which describes the outwards flowing radiation. This is joined smoothly to the vacuum Schwarzschild geometry at the boundary of the radiation zone.

to the body having a charge. In the case of a normal star, the effect of radiation on the overall exterior metric could be considered negligible when effects such as rotation, magnetic fields and so on, are considered which cause perturbations from spherical symmetry. However, the radiation effects would be relevant during the later stages of gravitational collapse when the star would be throwing away considerable mass in the form of radiation or when abundant supply of neutrinos is radiated from a collapsing supernova core (see for example, Kahana, Baron and Cooperstein, 1984).

Such a non-static distribution as the radiating star would then be surrounded by an ever-expanding zone of radiation. One could treat this radiating system, together with its radiation, as forming an isolated system in an otherwise empty, asymptotically flat universe. Then, beyond the zone of pure radiation, the space-time is described by the empty Schwarzschild solution (Fig. 19).

One is thus looking here for a spherically symmetric solution to Einstein equations $G_{ij} = 8\pi T_{ij}$ with the geometrical optics type stress-energy tensor for the radiation with form

$$T_{ij} = \sigma k_i k_j, \tag{3.60}$$

where k_i is a null vector radially directed outwards. The metric is best given in the null coordinates (u, r, θ, ϕ):

$$ds^2 = -\left(1 - \frac{2m(u)}{r}\right) du^2 - 2dudr + r^2 d\Omega^2, \tag{3.61}$$

with $m(u)$ being an arbitrary non-increasing function of the retarded time u. (In Section 6.4, where we shall be concerned with the application of Vaidya space-times to examine the cosmic censorship hypothesis, we will consider imploding radiation shells, rather than the outgoing case considered here. Then, the function m will be taken to be non-decreasing and the advanced null coordinate $t + r$ will be used.) The above gives the Vaidya metric in the radiation zone, which is to be matched by the interior metric of the radiating body at the boundary of the star and is matched by the Schwarzschild metric in the exterior.

In the form (3.60) for the energy-momentum tensor, σ is defined to be the energy density of radiation as measured locally by an observer with a four-velocity vector v^i. Thus, σ is the energy flux as well as energy density measured in this frame,

$$\sigma \equiv T_{ij} v^i v^j, \tag{3.62}$$

with $v^i v_i = -1$. Working out the connection coefficients from eqn (3.61), the Ricci tensor in null coordinates is given by

$$R_{ij} = -\frac{2}{r^2} \frac{dm(u)}{du} \delta^0{}_i \delta^0{}_j. \tag{3.63}$$

This implies that the Ricci scalar $R^i_i = R = 0$ and hence the Einstein equations give

$$T_{ij} = -\frac{1}{4\pi r^2} \frac{dm(u)}{du} \delta^0{}_i \delta^0{}_j, \tag{3.64}$$

which is the energy-momentum tensor of a radiating field in the geometric optics form. From eqns (3.62) and (3.64) we get

$$\sigma = -\frac{1}{4\pi r^2} \frac{dm(u)}{du}, \tag{3.65}$$

which is the expression for the energy density of radiation.

In the case when $m(u) = \text{const.}$ the relationship of the null coordinates in eqn (3.61) with the Schwarzschild coordinates (t, r, θ, ϕ) is not difficult to see. In such a case, one can use the transformation given by Finkelstein (1958) to diagonalize eqn (3.61)

$$u = T - r - 2m \log(r - 2m), \tag{3.66}$$

which gives the Schwarzschild metric in (T, r, θ, ϕ) coordinate system.

The energy flux from the star, as seen by an outside observer, was computed by Lindquist, Schwartz and Misner (1965) by considering only radially moving observers. As pointed out above, σ is the energy flux measured in a local frame. If $U \equiv v^r = dr/d\tau$ for a radially moving observer with $v^i v_i = -1, v^\theta = 0$ and $v^\phi = 0$, then from eqns (3.64) and (3.65) and denoting the energy density by q, one gets

$$q = -\frac{1}{4\pi r^2}\frac{dm}{du}(\gamma + U)^{-2}. \tag{3.67}$$

Since q must be positive, being energy density, it follows from the above that $dm/du \le 0$. For an observer at rest at infinity, the total luminosity is given by

$$L_\infty(u) = \lim_{r\to\infty, U=0} 4\pi r^2 q = -\frac{dm}{du}, \tag{3.68}$$

that is, it is the negative rate of change of mass of the radiating body.

The surface $r = 2m(u)$ has many interesting properties as pointed out by Lindquist, Schwartz and Misner (1965). Unlike the Schwarzschild case, where $r = 2m$ is a null hypersurface, for the Vaidya radiating star metric, this is a spacelike hypersurface. The induced metric on this hypersurface is given by

$$ds^2 \mid_{r=2m(u)} = -2\left(\frac{dm}{du}\right) du^2 + r^2 d\Omega^2. \tag{3.69}$$

This induced metric has the signature $(+, +, +)$ since $dm/du < 0$. As a result, the position of light cones on this surface is such that for all timelike vectors in the forward light cone at all points on this surface, we have $dr/du > 0$. Thus, no timelike trajectory from the outside region $r > 2m(u)$ can come and cross this surface to enter inside the $r < 2m(u)$ region.

The Vaidya solution discussed above is of type D in the Petrov classification of space-times and possesses a normal shear-free congruence with non-zero expansion. Kinnersley (1969) obtained a more general solution in the same class which describes the gravitational field of a radiating point particle which is in accelerated motion.

3.6 Robertson–Walker models

The Robertson–Walker models (Robertson, 1933) provide a major application of Einstein's general theory of relativity in cosmology. The discovery of Friedmann solutions (Friedmann, 1922) within the framework of homogeneous and isotropic universe models allowed the cosmological considerations to be treated in a mathematical manner, which was a subject so far dominated by largely speculative ideas about the overall structure of the universe. The major assumptions used in arriving at the Robertson–Walker geometry are the large scale homogeneity and isotropy of the universe. The homogeneity in space means that the universe is roughly the same at all spatial points and that the matter is uniformly distributed all over the space. This is an assumption difficult to check. Even though the universe is clearly inhomogeneous at the local scales of stars and star clusters, it is generally argued that an overall homogeneity will be achieved only at a large enough scale in a statistical sense only. It is possible to have observational tests on the assumption of isotropy, that is, the universe must be the same in all directions. One could check the distribution of galaxies in the different directions together with their apparent magnitudes and red-shifts, and also the distribution of radio sources similarly. Such observations are again interpreted frequently as providing an evidence for isotropic distribution of matter in the universe from our vantage point. Again, the observed microwave background radiation appears to be isotropic to a high degree of approximation in all directions. Then, if this radiation is of cosmic origin, it would imply that the perturbations from overall isotropy should not be very large on our past light cone to which all our observations are confined.

If we assume now the isotropy of the universe and combine it with the assumption of the *cosmological principle*, which is generally given by the statement that we do not occupy any special position in the universe, then the assumption of isotropy can be extrapolated to hold at all points of the universe. In such a case, it is then possible to deduce the homogeneity of the universe from the above assumption of isotropy at all points. This is what is meant when it is stated sometimes that the isotropy of the universe implies its homogeneity of mass distribution.

In such a case, the universe is spherically symmetric around all its points as opposed to the situation of the asymptotically flat space-times such as the Schwarzschild, where there is a spherical symmetry around the center. In terms of the mathematical model of space-time, the homogeneity assumption could be interpreted as saying that the space-time is a stockpile of spacelike hypersurfaces, each defining a constant value of time and given any two points x and y of one of these hypersurfaces, there is an isometry of the metric tensor g which takes x to y. Thus, topologically $M = \Sigma \times \mathbb{R}$, where Σ is a three-dimensional spacelike hypersurface and also the space-time is globally hyperbolic (Section 4.4) in the sense that all the non-spacelike curves in M must intersect Σ once and only once either in the future or in past. Then, given the initial data on any spacelike surface, the entire future and past evolution of the universe is completely fixed by the Einstein equations as we see in the Robertson–Walker case.

Obviously, it is actually not quite clear at the moment at what scale

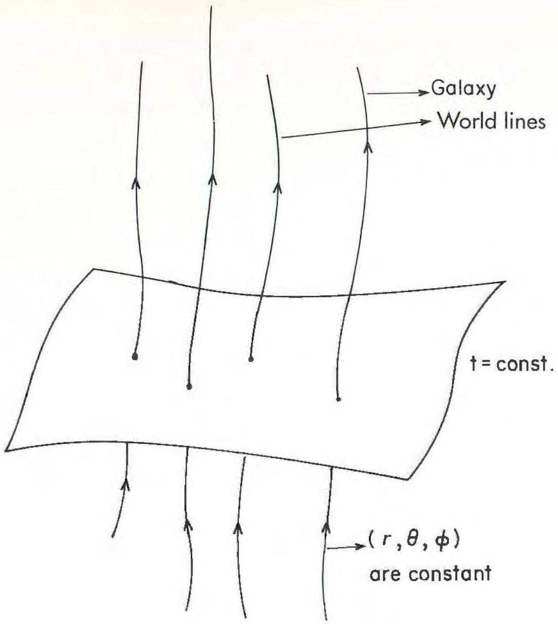

Fig. 20 Weyl's postulate. Galaxies, treated as point particles move along timelike geodesics orthogonal to the spacelike surfaces of constant time. The space coordinates (r, θ, ϕ) of each such particle are constant.

in the universe the homogeneity assumption will be realized. Observationally, there are local irregularities such as the stars and galaxies. In fact, observations tell us that such a homogeneity does not materialize even at the scale of the clusters of galaxies. The assumption thus means that the homogeneity of the universe is realized at a sufficiently large, but as yet undetermined, scale. (We will discuss this issue further in Chapter 8.)

In such an approximation clearly the galaxies, or even the clusters of galaxies are to be taken as points while modelling the universe and writing its line element depicting the global geometry. According to the *Weyl postulate*, the galaxies (or the clusters of galaxies) are assumed to be moving along timelike geodesics along which the space coordinates of the galaxy remain constant (Fig. 20). The time coordinate could then be chosen to be the proper time along such a geodesic world line. With the assumptions of homogeneity and isotropy as stated above and choosing the comoving coordinates, the metric for the space-time can be shown to have the following form in the (t, r, θ, ϕ) coordinates:

$$ds^2 = -dt^2 + S^2(t)\left[\frac{dr^2}{1-kr^2} + r^2(d\theta^2 + \sin^2\theta\, d\phi^2)\right]. \tag{3.70}$$

This is called the *Robertson–Walker line element*, where k is a constant which denotes the spatial curvature of the three-space and could be normalized to one of the values $+1$, 0, or -1. These cases correspond respectively to a three-space of constant positive curvature, a flat space, or a space of negative constant curvature. When k is zero, the space-time is called an *Einstein–deSitter universe.* The cases $k = +1$ and $k = -1$ are known as the *closed* and *open Friedmann models* respectively.

Assuming the matter content of the universe to be in the form of a perfect fluid given by

$$T_{ij} = (\rho + p)V_i V_j + p g_{ij},$$

the conservation equation $T^{ij}{}_{;j} = 0$ gives

$$\dot{\rho} + 3(\rho + p)\frac{\dot{S}}{S} = 0. \tag{3.71}$$

Here ρ is the density and p the pressure of the fluid. Next, one could solve the Einstein equations for the homogeneous and isotropic metric (3.70) to obtain the following two equations:

$$\frac{3\ddot{S}}{S} + 4\pi(\rho + 3p) = 0, \tag{3.72}$$

$$\frac{3\dot{S}^2}{S^2} - \left(8\pi\rho - \frac{3k}{S^2}\right) = 0. \tag{3.73}$$

(We assume here that the cosmological constant Λ is zero. If $\Lambda \neq 0$ there will be some qualitative changes in the picture of evolution of the universe presented here. We will discuss such possibilities in Chapter 8.) The following implications are immediate from the above equations. Firstly, if $\rho > 0$ and $p \geq 0$ then $\ddot{S}$ must be negative. Hence, $\dot{S} \neq$ const. and we must have $\dot{S} > 0$ or $\dot{S} < 0$, that is, the universe must be either expanding or contracting. This is a definite prediction arrived at by applying the general theory of relativity in the cosmological framework. The observations by Hubble of the red-shifts of the galaxies were interpreted by him as implying that all of them are receding from us with a velocity proportional to their distances from us and so the universe is expanding. This could be taken as a verification of the general theory of relativity as coming from the cosmological observations.

Next, since the universe is expanding and $\dot{S} > 0$, that implies from the above equations that $\ddot{S} < 0$. So, $\dot{S}$ is a decreasing function and at earlier times the universe must be expanding at a faster rate as compared to its present rate of expansion. Even if it expanded at a constant rate at all the times in the past, equal to its present expansion rate defined by,

$$\left(\frac{\dot{S}}{S}\right)_{t=t_0} \equiv H_0, \tag{3.74}$$

we must have $S = 0$ at a time

$$t = (H_0)^{-1}. \tag{3.75}$$

Since the expansion was faster in the past, actually $S = 0$ is realized earlier than the time given above. Thus, H_0^{-1} implies a global upper limit for the age of any type of Friedmann models. The quantity H_0 is called the *Hubble constant* and at any given epoch it measures the rate of expansion of the universe. Observations on the recession of galaxies as measured from observed redshifts imply that H_0 has a value somewhere in the range of 50 to 100 $\mathrm{km\,s^{-1}Mpc^{-1}}$. As a result the value of the above upper limit to the age of the universe lies around 10^{10} years with uncertainty of a factor of about two.

At the epoch $S = 0$, the entire three-space shrinks to zero size and the densities and curvatures grow to infinity. Thus, according to the Friedmann–Robertson–Walker models, the universe has an all encompassing space-time singularity at a finite time in the past. This curvature singularity is called the *big bang*. It was thought for quite some time that such a singularity and the singularities such as those arising at $r = 0$ in the Schwarzschild models are only due to the high degree of symmetry assumed for these models, such as the homogeneity and isotropy assumptions or the spherical symmetry. However, it was shown by the singularity theorems that such singularities are in fact a generic feature of the general theory of relativity and will exist in a space-time provided certain very general conditions including the positivity of energy and causality are satisfied. We will discuss such singularities in some detail in Chapter 5. It should be noted that there is a basic qualitative difference between the Schwarzschild singularity and that occurring in the Friedmann models discussed above. The Schwarzschild singularity could be the final result of a gravitationally collapsing configuration such as a massive star. However, the big bang singularity must be interpreted as the catastrophic event from which the entire universe emerged and where the usual concepts of space and time fully break down. The three-space here vanishes to a zero size and it makes no sense to ask what was there before this singularity.

If the matter content of the universe has a vanishing pressure $p = 0$, then we have a dust-filled universe. The solutions of Einstein equations for dust filled universes with $k = +1$ were given by Friedmann (1922). In general, the models with $k = \pm 1$ or $k = 0$ are called the *Friedmann models*. In order to examine the evolution of these models we note that the pressure is generally contributed by factors such as peculiar motions of galaxies, radiations present in the universe, magnetic fields, cosmic rays and so on. At the present epoch in the universe, one could take $p = 0$ as a good approximation when compared with the overall density ρ. In that case, the conservation equation (3.71) integrates to,

$$\rho S^3 = \text{const.} = C, \tag{3.76}$$

which provides the conservation of the rest mass. On the other hand, for the radiation, such as the microwave background radiation, the equation of state will be $p = \rho/3$, and the integration again gives

$$\rho S^4 = \text{const.} = C_1. \tag{3.77}$$

It follows that if the microwave background has a global origin such as the big bang singularity, then its density in the past will grow faster as compared to that of matter. So, even though the universe is matter dominated at the moment, it will become radiation dominated in the past and in the early phases soon after originating from the big bang singularity. The evolution of the radiation density in this phase is given by eqn (3.77).

The dynamical evolution of dust-filled Friedmann models is determined by the value of the parameter k (Fig. 21). One could substitute eqn (3.76) in the field equation (3.73), then given the integral of the conservation equation, the field equation involving $\ddot{S}$ becomes redundant and we need to integrate only the following obtained by the substitution above

$$\dot{S}^2 = \frac{8\pi C}{3}\frac{1}{S} - k. \tag{3.78}$$

For the case $k = 0$ of flat spatial sections which are non-compact and infinite in extent, the integral is given by

$$S = (6\pi C)^{1/3} t^{2/3}. \tag{3.79}$$

These spatial sections (and the universe) originate from the big bang singularity in the past and then expand forever in time with increasing S. For the cases $k = \pm 1$ the integration is best given in the parametric form as below. For $k = -1$, the spatial sections are hyperboloids of constant

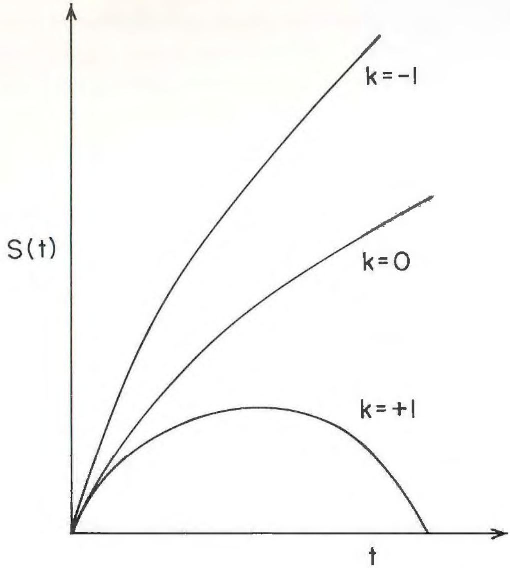

Fig. 21 The dynamical evolution of the dust-filled Friedmann models.

negative curvature and they are non-compact and infinite in extent. The solution is given in the parametric form by

$$S = \frac{4\pi C}{3}(\cosh\eta - 1), \qquad t = \frac{4\pi C}{3}(\sinh\eta - \eta). \tag{3.80}$$

The universe again originates in the big bang and continues to expand forever in time. In the case $k = +1$ the solution is

$$S = \frac{4\pi C}{3}(1 - \cos\eta), \qquad t = \frac{4\pi C}{3}(\eta - \sin\eta). \tag{3.81}$$

The spatial sections are now compact three-spheres with constant positive curvature and the radius at a time t is given by the scale parameter S. Here the expansion of the universe from the singularity reaches a maximum and then the scale factor again contracts to zero value ending in a singularity again. The equation for $S(t)$ is a cycloid. We note that in each of the cases discussed above, the present age of the universe will be less than the value H_0^{-1}.

The existence of a strong curvature big bang singularity in the past as indicated by the Friedmann models imply the existence of a very hot, dense and radiation dominated region in the very early phases of evolution of the universe. In such a phase, a copious production of elementary particles such as neutrinos could have taken place, which would then expand with the universe. If such particles have a tiny mass, they could constitute a substantial fraction of the total mass-energy density of the universe. In Chapter 8, we will discuss some such particle varieties which could have been produced in the early universe and also discuss general mass upper limits obtainable for such particles.

In the beginning of this section, we stated the cosmological principle as the requirement that the spacelike hypersurfaces of constant time are homogeneous and isotropic subspaces of the space-time. This leads to the determination of metric as the Robertson–Walker line element for the cosmological space-time and physically it means that there is no preferred position or a preferred direction in the universe for the observer. Often, a weaker version and also a stronger version of the above cosmological principle are invoked in cosmology, which we discuss below briefly. The *weak cosmological principle* states that the spacelike surfaces of simultaneity are homogeneous, that is, such surfaces admit three independent spacelike Killing vectors at any given point. Physically, for all fundamental observers on a given surface the state of the universe is the same and there is no preferred position in the universe. However, there is no assumption of global isotropy of the matter distribution imposed now. An example of exact solutions of Einstein equations obeying this weak principle is given by the Bianchi cosmological models, which are spatially homogeneous but anisotropic. This would permit rotation and shear in the motion of galaxies, which are often considered to be physically important features which one would like to be incorporated in cosmological considerations.

On the other hand, Bondi and Gold (1948) argued for a *perfect cosmological principle*, which is a stronger version of the cosmological principle and requires that in addition to the homogeneity and isotropy of the spacelike surfaces, the universe must look the same to all fundamental observers at all times as well. Their argument was, all the observations in cosmology are made by receiving the light rays which come from the past light cone of the observer, which were emitted a long time ago and hence one must assume the constancy of the laws of physics over this entire time interval. Effectively, their assumption means that there is no preferred position, preferred direction, or any preferred epoch of time in the universe. This amounts to assuming the existence of a timelike Killing vector field in the space-time and this is possible only if in the Robertson–Walker line element either one has $S(t) = \text{const.}$ or $S(t) = \exp(Ht)$ with $H \neq 0$ and the spatial curvature $k = 0$. Such cosmological models are called the *steady state models* and were developed by Bondi and Gold (1948), Hoyle (1948), and Hoyle and Narlikar (1964) through different approaches (for details we refer to Weinberg (1972) or Narlikar (1983)). The line element for the universe is fixed in this case uniquely as

$$ds^2 = -dt^2 + e^{2Ht}[dr^2 + r^2(d\theta^2 + \sin^2\theta\, d\phi^2)].$$

Here H could be positive, negative, or zero, and its value is to be fixed from observations. The value $H = 0$ is ruled out because it leads to a static universe with infinite sky temperature. Similarly, for $H < 0$ we have a contracting universe with radiation from distant light sources blue shifted, which again gives an infinite sky background. Thus H must be positive, corresponding to an expanding universe.

An important difference between the Friedmann universes discussed earlier and the steady state models is that the later have no singularity of infinite curvatures and density either in the past or in the future. This is due to the basic reason that in the steady state theory, the universe is the same as its present state at all times in the past and also in future. Thus, the density of the universe is constant at all epochs in a steady state universe. In order to match this with the expansion of universe, the creation of matter at a constant rate is required in the steady state theory. However, considering the present density of the universe, which is of the order of $10^{-30}\mathrm{gm\,cm^{-3}}$, and the present expansion rate stated earlier, this creation rate turns out to be extremely small, of the order of $10^{-48}\mathrm{gm\,cm^{-3}\,s^{-1}}$. It would appear that there is no possibility to detect the same with presently available instruments.

4

CAUSALITY AND SPACE-TIME TOPOLOGY

The condition that no material particle signals can travel faster than the velocity of light fixes the causal structure for Minkowski space-time. In general relativity also, which uses the framework of a general space-time manifold, locally the causality relations are the same as in the Minkowski space-time. However, globally there could be important differences in the causal structure due to a different space-time topology, strong gravitational fields and so on. This causal structure of space-time has been studied in detail, especially in view of the occurrence of space-time singularities in gravitational collapse and cosmology (see for example, Penrose, 1972; Hawking and Ellis, 1973; Geroch, 1970b).

The purpose of this chapter is to investigate further this causality structure of space-time and its relationship with the space-time topology. The known results and definitions required for later chapters are also reviewed here and several results are established which are of intrinsic interest for the global aspects of space-times and which emphasize the close interplay between the causal structure and topology.

In Section 4.1 we set up basic causal structure ideas and definitions. Topological properties of several space-time sets such as chronological futures and pasts and various causality conditions are discussed in Section 4.2 in order to arrive at a unified causality statement for a reasonable model of space-time. Several properties of reflecting and causally continuous space-times are established in Section 4.3 and global hyperbolicity is studied in Section 4.4 with a particular reference to the maximal length property of non-spacelike curves between chronologically related events in such a space-time. Causality preserving space-time mappings are discussed in Section 4.5 and Section 4.6 introduces and develops the tool of causal functions in a space-time which provides an integrated characterization for many ideas in causal structure. In Section 4.7 we study asymptotically flat space-times and it is shown there by an explicit analysis of the Schwarzschild and Kerr-Newman space-times as to how the light cone cuts of null infinity are related to the interior causal structure and the metric when the asymptotic flatness is respected, admitting a well-defined null infinity.

4.1 Causal relations

As pointed out in Chapter 2, a *space-time* (M, g) is taken here to be a connected, C^∞ Hausdorff differentiable manifold which is paracompact and admits a Lorentzian metric g of signature $(-,+,+,+)$. Further, the space-time is also assumed to be space and time-oriented. The considerations in later chapters will be mostly confined to space-times of dimension four but results in the present chapter are valid for manifolds of arbitrary dimensions greater than or equal to two.

An event p *chronologically precedes* another event q, denoted by $p \ll q$, if there is a smooth future directed timelike curve from p to q. If such a curve is non-spacelike that is, either timelike or null, we say that p *causally precedes* q, or $p < q$. The *chronological future* $I^+(p)$ of p is the set of all points q such that $p \ll q$. The *chronological past* of p is defined dually. Thus we have

$$I^+(p) = \{q \in M \mid p \ll q\}, \tag{4.1}$$

$$I^-(p) = \{q \in M \mid q \ll p\}. \tag{4.2}$$

The *causal future* (*past*) for p can be defined similarly

$$J^+(p) = \{q \in M \mid p < q\}, \tag{4.3}$$

$$J^-(p) = \{q \in M \mid q < p\}. \tag{4.4}$$

The relations $\ll$ and $<$ are seen to be transitive. Further, $p \ll q$ and $q < r$ or $p < q$ and $q \ll r$ implies $p \ll q$ (Penrose, 1972). It is seen from this that

$$\overline{I^+(p)} = \overline{J^+(p)}, \tag{4.5}$$

and also

$$\dot{I}^+(p) = \dot{J}^+(p), \tag{4.6}$$

where for a set A, $\bar{A}$ is the closure of A and $\dot{A}$ denotes the topological boundary. The chronological(causal) future of any set $S \subset M$ is defined in a similar way:

$$I^+(S) = \bigcup_{p \in S} I^+(p), \tag{4.7}$$

$$J^+(S) = \bigcup_{p \in S} J^+(p). \tag{4.8}$$

The chronological (causal) pasts for subsets of space-time are similarly defined. Various such dual definitions and results will often be taken here for granted. Now, suppose there is a future directed timelike curve from p to q. There is a local region containing q which the timelike curve must enter, in which special relativity is valid. Thus, there is a neighbourhood N of q such that any point of N can be reached by a future directed timelike curve from p. It follows that

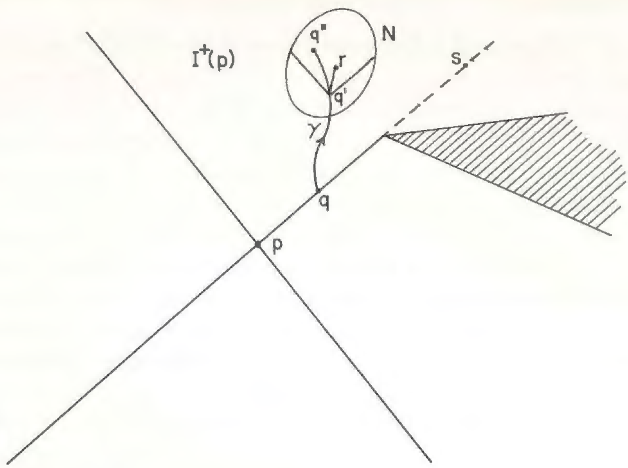

Fig. 22 Removal of a closed set from the space-time gives a causal future $J^+(p)$ which is not closed. Events p and s are not causally connected, though $s \in \overline{I^+(p)} = \overline{J^+(p)}$. There is a null geodesic from p to q and a timelike curve γ from q to r. Then the corner at q can be smoothed to give a smooth timelike curve from p to r. The set $I^+(p)$ is open because for $r \in I^+(p)$, one could choose q' on γ sufficiently near to r in a convex normal neighbourhood of r, then the set of points $q'' \in I^+(q')$ forms an open neighbourhood of r in $I^+(p)$.

Lemma 4.1 (Penrose, 1972). For any event $p \in M$, the sets $I^+(p)$ and $I^-(p)$ are open in M.

The above lemma also implies that the sets $I^\pm(S)$ are open, being union of open sets in M. However, the sets $J^\pm(p)$ are neither open nor closed in general. (See, for example Fig. 22, where removing a closed set from the Minkowski space-time gives a causal future $J^+(S)$ of a closed set S which is not closed.) Thus, all the points in the boundary of $J^+(S)$ are not necessarily connected to a point in S by a null geodesic generator.

In Minkowski space-time, $I^+(p)$ is the set of those points which are reached by future directed timelike geodesics from p and the boundary of this set is generated by null geodesics from p. As seen above, this is not true for an arbitrary space-time in general, but locally this property is still valid as shown by the Proposition 4.5.1 of Hawking and Ellis (1973).

Proposition 4.2. Let (M, g) be a space-time manifold and N be a convex normal neighbourhood of $p \in M$. Then, for any $q \in N \cap I^+(p)$, there is a timelike geodesic from p to q in N and the boundary of $I^+(p)$ in N is generated by future directed null geodesics from p in N.

To derive properties of more general boundaries, we need the definition of a future set: A set F of M is called a *future set* if $F = I^+(S)$ for some subset S of M. An equivalent criterion that is we must have, $I^+(F) \subset F$. Past sets are defined dually. Clearly, future sets are open in M being union of $I^+(p)$ for all $p \in S$.

To get an idea of properties of future sets, we note that if F is a future set, $\overline{F}$ is the set of all points x such that $I^+(x) \subset F$. To see this, suppose x is such that $I^+(x) \subset F$. Then one can construct a sequence in $I^+(x)$, and hence in F, with the limit point x. Thus $x \in \overline{F}$. Conversely, if $x \in \overline{F}$, then take $y \in I^+(x)$. This gives $x \in I^-(y)$, which contains an open neighbourhood of x. Then, this neighbourhood contains points of F, that is, $y \in I^+(F)$, which implies $I^+(x) \subset F$.

Again, the boundary of a future set F is made of all events x such that $I^+(x) \subset F$ but $x \notin F$. If $x \in \dot{F}$ then clearly $x \notin F$ as F is an open set. But $x \in \overline{F}$ and as seen above $I^+(x) \subset F$. Conversely, let $x \notin F$ but $I^+(x) \subset F$. Then one can clearly construct a sequence in $I^+(x)$ converging to x which implies $x \in \overline{F}$. Thus x must be in the boundary of F.

A set S is called *achronal* if no two points of S are timelike related, that is, $I^+(S) \cap S = \emptyset$. If F is a future set, the following result shows that the boundary of F is a well-behaved achronal manifold (Hawking and Ellis, 1973)

Proposition 4.3. Let F be a future set, then the boundary of F is a closed, achronal C^0 manifold which is a three-dimensional embedded hypersurface.

The following result (Penrose, 1972) then shows that the achronal boundary of a future set F is always generated by null geodesics which are either past endless or always have a past end point on $\overline{F}$.

Proposition 4.4. Let $S \subset M$ and $p \in \dot{I}^+(S) - \overline{S}$. Then there exists a null geodesic contained in the boundary of $I^+(S)$ with future end point p and which is either past-endless or has a past endpoint on $\overline{S}$.

From the above it follows now that if $q \in J^+(p) - I^+(p)$ then any non-spacelike curve joining p and q must be a null geodesic, and that the boundary of $I^+(p)$ or $J^+(p)$ is generated by null geodesics which have either a past end point at p or are past endless.

Causal relations in a space-time are defined by the existence of smooth non-spacelike curves between pairs of events. It is, however, useful to extend this to define causality by means of continuous curves. This is done by requiring that pairs of points on a curve are locally joined by a smooth timelike or causal curve. To be precise, a *continuous* curve λ is called a *future directed timelike* (*or non-spacelike*) if each $x \in \lambda$ is contained in a convex normal neighbourhood N such that if $\lambda(t_1), \lambda(t_2) \in N$ with $t_1 < t_2$, then there is a smooth future directed timelike (non-spacelike) curve in

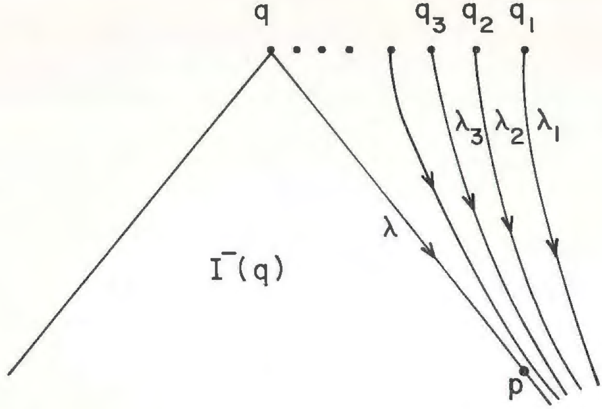

Fig. 23 Any neighbourhood of the point p intersects an infinitely many non-spacelike curves of the sequence $\{\lambda_n\}$. The curve λ is a limit curve.

N from $\lambda(t_1)$ to $\lambda(t_2)$. Such curves are regarded as equivalent under a one-to-one continuous reparametrization.

It is useful to introduce at this stage the notion of future and past inextendible non-spacelike curves, which are effectively the trajectories which have no future or past end points. Let λ be a non-spacelike curve. Then $p \in M$ is called a *future end point* of λ if for every neighbourhood N of p, there exists a value of the curve parameter t' such that for all $t > t'$ we have $\lambda(t) \in N$. The past end point is defined similarly. It is clear that if λ has an end point, it must be unique because M is Hausdorff. The curve λ is called *future or past inextendible* if it has no future or past end point respectively in M. An inextendible causal curve might be running off to infinity, or it might end up in a space-time singularity, or it could enter a compact set in which it could be trapped to go round and round forever.

Next, let $\{\lambda_n\}$ be a sequence of non-spacelike curves. A point $x \in M$ is called a limit point of the sequence $\{\lambda_n\}$ if every open neighbourhood of x intersects infinitely many λ_n. A curve λ is called a limit curve of $\{\lambda_n\}$ if for any point $p \in \lambda$ and any open neighbourhood U of p, there is a subsequence $\{\lambda'_n\}$ such that for sufficiently large n all $\{\lambda'_n\}$ intersect U (Fig. 23). It follows that if λ is a limit curve, each $p \in \lambda$ is a limit point of $\{\lambda_n\}$. The following result plays an important role in establishing the existence of space-time singularities to be discussed in the next chapter.

Proposition 4.5. Suppose $\{\lambda_n\}$ is a sequence of future inextendible non-spacelike curves with a limit point p. Then there is a future inextendible non-spacelike curve λ through p which is a limit curve of $\{\lambda_n\}$.

Finally, we note here that the study of casual relations in a space-time (M, g) is equivalent to that of the conformal geometry of M. Let (M, g) be the physical space-time and consider the set of all conformal metrics $\bar{g}$ where $\bar{g} = \Omega^2 g$, with Ω being a non-zero C^r function. Then if $p \ll q$ or $p < q$ in (M, g), the same relation is preserved in $(M, \bar{g})$ for any conformal metric $\bar{g}$. Thus casual relationships are invariant under a conformal transformation of the metric. However, non-spacelike geodesics in (M, g) will no longer be geodesics in $(M, \bar{g})$ unless they are null. The null geodesics will be conformally invariant up to a reparametrization of the affine parameter along the curve. Thus, specifying causal relations in M fixes the space-time metric up to a conformal factor. Let p be an event in M and let N be a convex normal neighbourhood of p. One could introduce Minkowskian coordinates $\{x^i\}$ in N in which case it follows from Proposition 4.2 that the set of events in N which are causally connected to p satisfy

$$-(x^0)^2 + (x^1)^2 + (x^2)^2 + (x^3)^2 \leq 0. \tag{4.9}$$

The boundary of these points define the null cone in the tangent space T_p. Suppose now W and Z are any two non-null vectors in T_p; then we can write

$$g(W, Z) = \tfrac{1}{2}\left[g(W + Z, W + Z) - g(W, W) - g(Z, Z)\right]. \tag{4.10}$$

Now, if $X, Y \in T_p$ are a timelike and a spacelike vector respectively, then the equation

$$g(X + \lambda Y, X + \lambda Y) = g(X, X) + 2\lambda g(X, Y) + \lambda^2 g(Y, Y) = 0 \tag{4.11}$$

has two distinct roots λ_1 and λ_2 as $g(X, X) < 0$ and $g(Y, Y) > 0$. The knowledge of null cone then implies that λ_1, λ_2 can be determined in principle. But $\lambda_1 \lambda_2 = g(X, X)/g(Y, Y)$. Hence the null cone gives the ratio of the magnitudes of a timelike and a spacelike vector. Thus, each term in eqn (4.11) is determined up to a factor so $g(W, Z)$ is determined up to a factor.

Finally, we have introduced the assumption of *time orientability*, on M, that is, at all points p it is possible to make a continuous designation of future and past of p in a consistent manner throughout the manifold. Thus, at every point the timelike vectors are divided into a future directed class and a past directed class continuously over the entire space-time. In special relativity, one part of the null cone contains future directed timelike vectors and the other part contains past directed timelike vectors. Such a choice may not always be possible for a general space-time globally. This is illustrated by the example of the Möbius strip discussed in Section 2.2, when given an indefinite metric. In that case, a consideration of light cones along x-axis shows that when one comes back to the same point after going along a closed curve, the future and past directions are reversed. In such a case an arrow of time is not clearly defined on the space-time and the notions of going forward or backward in time are not available in a consistent manner.

The time orientability of M is equivalent to the condition that the space-time admits a continuous, nowhere-vanishing timelike vector field V which is then used to separate the non-spacelike vectors at each point into the future and past directed classes. When a smooth, non-vanishing timelike vector field t^a exists globally on M, one could arbitrarily label as future or past the cone in which t^a lies, thus providing a consistent time-orientation throughout M. Conversely, suppose a continuous choice of time-orientation is given globally on M in a consistent manner. Then, the paracompactness of M implies that there is a smooth, positive definite Riemannian metric K_{ab} defined on M. Now, consider the set of timelike vectors v^a such that $K_{ab}v^a v^b = 1$ and v^a is future directed. This being a compact subset of the tangent space, there exists t^a such that $g_{ab}t^a t^b$ is minimum over this set. This t^a varies smoothly over M providing the desired smooth vector field.

4.2 Causality conditions

The local causality principle for a space-time implies that over small regions of space and time the causal structure is the same as in the special theory of relativity. However, as soon as one leaves the local domain, global pathological features may show up in the space-time such as the violation of time orientation, possible non-Hausdorff nature or non-paracompactness, having disconnected components of space-time, and so on. Such pathologies are to be ruled out by means of 'reasonable' topological assumptions only. In particular, one would like to ensure that the space-time is causally well-behaved. This can be done by means of introducing various causality conditions such as non-occurrence of closed timelike or non-spacelike curves (causality), and stability of this condition under small perturbations in the metric (stable causality). In fact, Carter (1971) has pointed out that there is an infinite hierarchy of such causality conditions for a space-time.

It would appear reasonable to demand that a physically realistic space-time should not allow either closed timelike or closed non-spacelike curves, because this would give rise to the phenomenon of entering one's own past. However, general relativity and Einstein's equations as such do not rule out such a possibility on their own. For example, Gödel universe has closed timelike curves through each point of M. Again, global topology of M can cause closed timelike curves. For example, for the cylinder

$M = S^1 \times \mathbb{R}$, obtained from the Minkowski space-time by identifying $t = 0$ and $t = 1$ hypersurfaces with the metric given by $ds^2 = -dt^2 + dx^2$, the circles $x =$ const. are closed timelike curves. In fact, for all $p \in M$ here, $I^+(p) = I^-(p) = M$. One could dismiss the examples such as above as mere mathematical pathologies in the space-time topology; and the Gödel universe may be termed as unrealistic because it corresponds to a rotating model which does not correspond to the universe we observe. More difficult to rule out are the Kerr solutions of a spinning gravitational source which would contain closed timelike curves if the rotation is sufficiently fast. These could possibly represent the final fate of a massive collapsing star which is rotating. If such a star failed to get rid of enough spin during the process of collapse, it would give rise to a time machine in the space-time. Wormholes in a space-time representing the multiply connected nature of the topology of space could also give rise to closed timelike curves as indicated in Chapter 2. The physical significance and acceptability of such causality violations have been examined by Morris, Thorne and Yurtsever (1988), and by Friedman *et al.* (1990). It follows from these considerations that such wormhole space-times in fact do admit unique solutions to at least simple field equations, such as a single non-interacting scalar field. This shows that even though the causality violation is perceived as contradictory to the predictability requirement in the space-time, in general this need not be so in that there are wormhole space-times where this is not the case.

The situation of having a space-time with closed non-spacelike curves is avoided by requiring M to satisfy the *causality condition*, that is, (M, g) does not admit any closed timelike or null curves. A space-time M is said to be *chronological* when it admits no closed timelike curves, that is, $p \notin I^+(p)$ for all $p \in M$. When M admits no closed non-spacelike curves, it is said to be causal. If for $M = S^1 \times \mathbb{R}$ we choose the metric to be $ds^2 = dtdx$, then M is chronological but not causal. The circles $x =$ const. are null geodesics here. The causally well-behaved nature of M turns out to be closely related to the topological structure of M as shown by the following result.

Proposition 4.6. If M is chronological, M cannot be compact.

Proof. Suppose M is compact. The sets $\{I^+(p) \mid p \in M\}$ cover M. Now compactness implies that there exists a finite set of points $p_1, ..., p_n$ such that the set $I^+(p_1) \cup ... \cup I^+(p_n)$ covers M. Now, $p_1, ..., p_n$ must be in the cover implies $p_1 \gg p_{i_1}$ for some i_1 in $1, ..., n$. Thus we have $p_1 \gg p_{i_1} \gg <\gg p_{i_{n-1}}$ where we have exhausted all n-points. Hence $p_{i_k} \gg p_{i_k}$ for some i_k which violates chronology. □

We introduce here the notion of *chronological common future* for a space-time set which will be used later as well. For an open set $S \subset M$, this is defined as the largest open set $\uparrow S$ such that each of its events can be given a timelike signal from all events in S. That is,

$$\uparrow S = Int\,\{z \in M \mid \text{for all}\, s \in S, s \ll z\}\,. \tag{4.12}$$

Here Int denotes the topological interior of a set. The chronological common past $\downarrow S$ is defined dually. Alternatively, one can also define

$$\uparrow S = Int\left(\bigcap_{s\in S} I^{+}(s)\right). \tag{4.13}$$

The following result indicates that the causality violation at an event is connected in a way to the chronological common past of that event.

Proposition 4.7. The following are equivalent for a space-time M:
(a) M is chronological,
(b) $z \notin \uparrow I^{-}(z)$ for all $z \in M$,
(c) $z \notin I^{+}(z)$ for all $z \in M$.

Proof. The equivalence (a) and (c) is obvious. Now if $z \notin \uparrow I^{-}(z)$, then $z \notin I^{+}(z)$ also, since $I^{+}(z) \subset \uparrow I^{-}(z)$ is always true. So M is chronological.

Conversely, if there are no closed timelike loops then $z \notin I^{-}(z)$ for all $z \in M$ so $z \in \dot{I}^{-}(z)$ and every open set containing z will contain points of $I^{-}(z)$. Now, if $z \in \uparrow I^{-}(z)$, which is an open set, then there will be an open set N containing z such that $N \subset \uparrow I^{-}(z)$. So $\uparrow I^{-}(z)$, will contain points of $I^{-}(z)$. If p is such a point, then by definition of $\uparrow I^{-}(z)$, there is a future directed timelike curve from p to p and hence a closed timelike loop is created, which is a contradiction. □

Even though a space-time may be causal, it could be on the verge of violating causality. Consider (M, g) given by $ds^2 = dtdx + t^2 dx^2$. To see the behaviour of light rays, consider the null geodesics of this space-time which are given by $dt/dx = -t^2$. At $t = 0$ one has $dt/dx = 0$ and so one arm of the light cone will lie along the x-axis. In a situation like this, one would have a space-time (M, g) which is causal but a non-spacelike curve from p can enter arbitrary small neighbourhoods of p (Fig. 24). For distinct events p, q along x-axis, $I^{+}(p) = I^{+}(q)$. To avoid such a causal pathology, one could impose the *distinguishing condition* on M, namely, for all p, q in M, $I^{+}(p) = I^{+}(q)$ implies $p = q$ and $I^{-}(p) = I^{-}(q)$ implies $p = q$. A similar but stronger condition is *strong causality* (Penrose, 1972) which states that for all events $p \in M$, every neighbourhood of p contains a neighbourhood of p which no non-spacelike curve in M intersects more than once. If the

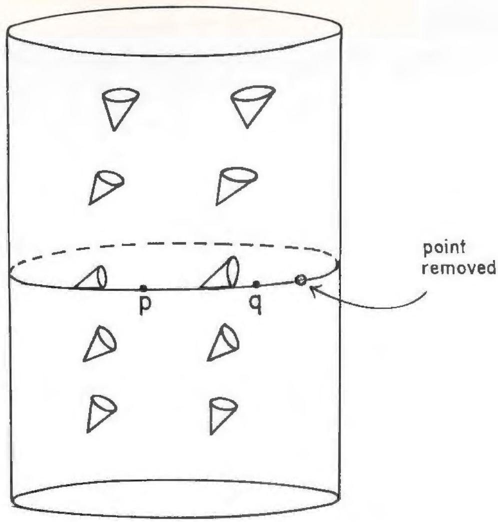

Fig. 24 The space-time here is causal because of the point removed, but it is not distinguishing because for different points p and q, $I^+(p) = I^+(q)$. The line through p is a null geodesic, and non-spacelike curves from p can come back to enter arbitrarily small neighbourhoods of p. There are no closed timelike or null curves in the space-time.

strong causality is violated at p then there are non-spacelike curves from neighbourhoods of p which come arbitrary close to intersecting themselves.

The relevance of strong causality condition is that it determines the topology of space-time. First, we note that the *Alexandrov topology* or the interval topology on M is defined by taking the collection $\{I^+(p) \cap I^-(q) \mid p, q \in M\}$ as the basis for topology on M. One can then show the following (Penrose, 1972).

Proposition 4.8. Let (M, g) be a strongly causal space-time. Then the manifold topology on M is the same as the Alexandrov topology.

Thus, for a strongly causal space-time, the causal relations determine the topological structure of M. It can also be shown further that the Alexandrov topology will be Hausdorff for a strongly causal space-time. Next, strong causality imposes a regularity on M in the sense that if γ is any future inextendible non-spacelike curve, then γ cannot be either totally or partially future imprisoned in any compact set S in the sense that γ will enter and remain within S or it will continually reenter S. Thus, for a strongly causal M, the future endless curve γ must go either to infinity or terminate in a space-time singularity, that is, it must go to the boundary

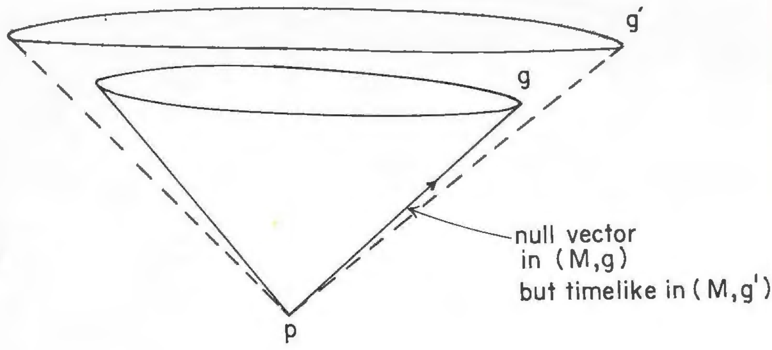

Fig. 25 The light cones are shown for $g' > g$. The null vectors in g become timelike in g'.

of space-time in either case.

Even though (M, g) could be strongly causal, it is possible to create examples of space-times where a small perturbation in the metric tensor components g_{ij} will give rise to closed timelike curves. Thus, a strongly causal space-time could still be on the verge of causality violation. Now, general relativity is supposed to be a classical approximation of some, as yet unknown quantum theory of gravity in which a precise measurement of the metric components at a single event would not be possible. Thus one would like the causality of space-time to be preserved under small perturbations in the metric. This is achieved by requiring the space-time to be *stably causal.* Let (M, g) be a space-time, then there exists another metric $g' > g$ such that there is no closed non-spacelike curve in g' if there was none in g. Here $g' > g$ means that every non-spacelike vector in g is always timelike in g' (Fig. 25). It is of course clear, by definition, that if $g' < g$ then g' will have no closed non-spacelike curves if g had none.

In order to understand the working and meaning of various causality conditions, and stable causality in particular, it is useful to study at this stage certain future sets in space-time. For example, the chronology condition on M is characterized in terms of future sets $I^+(p)$ as $p \notin I^+(p)$ for all $p \in M$; or $p \in I^+(q), q \in I^-(p)$ implies $p = q$. For this purpose, we introduce the sets *Seifert future* (Seifert, 1971) and *almost future* (Woodhouse, 1973) denoted by $J_S^+(x)$ and $A^+(x)$ as below.

$$J_S^+(x) = \bigcap_{g'>g} J^+(x, g'). \tag{4.14}$$

To define the almost future, the idea is to think of an event as the limit of a converging sequence of neighbourhoods. Thus, an event x *almost* *causally precedes* another event y, denoted by xAy, if for all $z \in I^-(x)$, $I^+(z) \supset I^+(y)$. It would be more reasonable physically to say that 'x almost causally precedes y' rather than saying 'there is a non-spacelike curve from x to y'. This is because if xAy then every neighbourhood of x in the manifold topology contains events which precede chronologically some events in any neighbourhood of y. Thus, no physical experiment could tell that x and y are not causally related. We now define

$$A^+(x) = \{y \in M \mid xAy\}. \tag{4.15}$$

It is clear that $y \in A^+(x)$ if and only if $x \in A^-(y)$.

Some properties of the sets $J_S^+(x)$ given by Seifert (1971) are as below.

Lemma 4.9. $J_S^+(x)$ is closed for all $x \in M$.

Lemma 4.10. Let $\{p_n\}$ be a sequence in $I^-(p)$ such that $p_n \to p$. Then,

$$\bigcap_{n=1}^{\infty} J_S^+(p_n) \subset J_S^+(p). \tag{4.16}$$

Lemma 4.11. A space-time M is stably causal if and only if J_S^+ is a partial ordering on M.

Since J_S^+ is reflexive and transitive by definition, this means that the antisymmetry of J_S^+ is equivalent to the stable causality of M. Thus, M will be stably causal if and only if $x \in J_S^+(y)$ and $y \in J_S^+(x)$ implies $x = y$ for all $x, y \in M$.

M is called *W-causal* if $x \in A^+(y)$ and $y \in A^+(x)$ implies $x = y$ for all $x, y \in M$. Similar to stable causality, the purpose of the above causality condition is also to take the metric perturbations into account. However, we will show that in general stable causality is a stronger condition between the two. Both these conditions, together with many other causality conditions, are shown to be equivalent in the next section when M is a reflecting space-time.

The following results give certain properties and a comparison between the almost future, chronological common future and the Seifert future.

Proposition 4.12. The almost future $A^+(x)$ is closed in the manifold topology for all $x \in M$.

Proof. Let p be a limit point of $A^+(x)$ in the manifold topology on M. Then there is a sequence $\{p_n\}$ in $A^+(x)$ such that $p_n \to p$. Let now $y \in I^+(p)$, then $p \in I^-(y)$, which is open, and hence a neighbourhood of p. So there exists a positive integer n_0 such that $p_n \in I^-(y)$ for all $n > n_0$. So $y \in I^+(p_n)$ for all $n > n_0$. Since $p_n \in A^+(x)$ for all n, this give

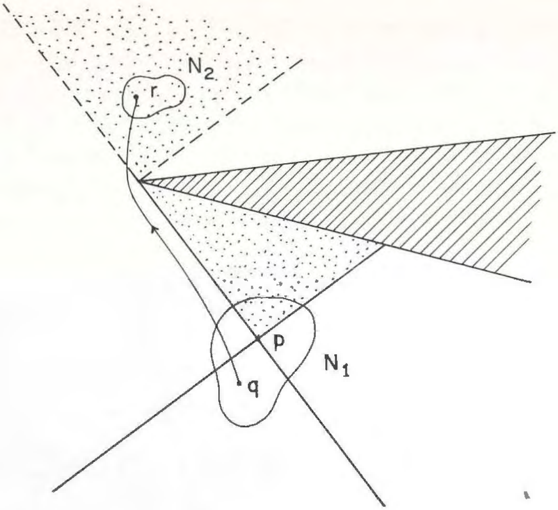

Fig. 26 Widening the light cones slightly gives rise to a sudden change in $I^+(p)$ to include the entire dotted region. The shaded region denotes a cut in the space-time. Here $r \in A^+(p)$ but r is not in $I^+(p)$. From every neighbourhood N_1 of p, there is a non-spacelike curve to every neighbourhood N_2 of r. Also, $r \in \uparrow I^-(p)$. Further, $I^-(r) \supset I^-(p)$, however, there is no intersection between $I^+(p)$ and $I^+(r)$.

$I^+(p_n) \subset I^+(z)$ for all $z \in I^-(x)$. So $y \in I^+(z)$ for all $z \in I^-(x)$. Thus, finally we have $I^+(p) \subset I^+(z)$ for all $z \in I^-(x)$ and so $p \in A^+(x)$. □

In Fig. 26, perturbing g slightly suddenly increases $J^+(x)$ to include the entire dotted region. Thus, $y \in A^+(x)$, and we also have $y \in \uparrow I^-(x)$. In general we have

Proposition 4.13. $Int[A^+(x)] = \uparrow I^-(x)$.

Proof. Let $p \in Int[A^+(x)]$. Then there is a simple region N containing p and contained within $A^+(x)$. Let $q \in I^-(p) \cap N$. Then $p \in I^+(q)$ and $q \in A^+(x)$. Therefore $I^+(q) \subset I^+(z)$ for all $z \in I^-(x)$. This gives $p \in I^+(z)$ for all $z \in I^-(x)$. In a similar manner one can see that any point of N belongs to $I^+(z)$ for all $z \in I^-(x)$. Hence $p \in \uparrow I^-(x)$.

Conversely, suppose $p \in \uparrow I^-(x)$. Since $\uparrow I^-(x)$ is an open set, there is a simple region N such that $p \in N \subset \uparrow I^-(x)$. So every point of N could be joined by a timelike curve from any point of $I^-(x)$. This gives $p \in N \subset A^+(x)$. Hence $p \in IntA^+(x)$. □

Thus, the almost future $A^+(x)$ is determined in some sense by the overall past of the point x, that is, $I^-(x)$. The following makes this more precise.

Proposition 4.14. Let $x \in M$ and $\{x_n\}$ be a sequence in $I^-(x)$ such that $x_n \to x$. Then

$$A^+(x) = \bigcap_{n=1}^{\infty} \overline{I^+(x_n)} = \bigcap_{n=1}^{\infty} \overline{J^+(x_n)}.$$

Proof. Let $p \in A^+(x)$. Then $I^+(p) \subset I^+(x_n)$ for all n. Now let $s \in \overline{I^+(p)}$. Then $I^+(s) \subset I^+(p) \subset I^+(x_n)$, so $s \in \overline{I^+(x_n)}$. This gives $p \in \overline{I^+(p)} \subset \overline{I^+(x_n)}$ for all n. Hence $p \in \cap_{n=1}^{\infty} \overline{I^+(x_n)}$ and so $A^+(x) \subset \cap_{n=1}^{\infty} \overline{I^+(x_n)}$.

Conversely, let $p \in \cap_{n=1}^{\infty} \overline{I^+(x_n)}$. So $p \in \overline{I^+(x_n)}$ for all n, implying that $I^+(p) \subset I^+(x_n)$ for all n. Now let $z \in I^-(x)$. Then $x \in I^+(z)$, and $I^+(z)$ being open, $x_n \in I^+(z)$ after certain stage. Then for such x_n, $I^+(z) \supset I^+(x_n) \supset I^+(p)$. So $p \in A^+(x)$. Since $\overline{I^+(x)} = \overline{J^+(x)}$ is always true, the equality $A^+(x) = \cap_{n=1}^{\infty} \overline{J^+(x_n)}$ follows. □

The relationship between the almost future and the Seifert future of an event in a general space-time is obtained in the following result.

Proposition 4.15. Let M be a space-time. Then for all $x \in M, A^+(x) \subset J_S^+(x)$.

Proof. By the definition of $J_S^+(x)$, clearly $I^+(x) \subset J_S^+(x)$. Now by Lemma 4.9, $J_S^+(x)$ is closed, hence $\overline{I^+(x)} \subset J_S^+(x)$ Let now $p \in A^+(x)$. Then $I^+(p) \subset I^+(z)$ for all $z \in I^-(x)$. Consider now a sequence $\{p_n\}$ in $I^-(x)$ converging to x. Then by Lemma 4.10 we have $\cap_{n=1}^{\infty} J_S^+(p_n) \subset J_S^+(x)$. Now $I^+(p) \subset I^+(z)$ for all $z \in I^-(x)$ implies that $p \in \overline{I^+(z)}$ for all $z \in I^-(x)$. So $p \in J_S^+(z)$ for all $z \in I^-(x)$ and so $p \in \cap_{n=1}^{\infty} J_S^+(p_n)$, since $p_n \in I^-(x)$ for all n. This gives $p \in J_S^+(x)$. □

It is now possible to see that W-causality of M is weaker than stable causality condition. Suppose M is stably causal and $x \in A^+(y)$ and $y \in A^+(x)$ holds for $x, y \in M$. Then the above implies $x \in J_S^+(y)$ and $y \in J_S^+(x)$. So by stable causality $x = y$ and hence M is W-causal. It is not difficult to see by means of a counter-example that the converse need not be true in general, that is, a W-causal space-time need not be stably causal.

Of the infinite hierarchy of causality requirements available for a space-time, stable causality may be considered to be the most relevant physically, which provides a unified criterion for basic causal regularity. Such a space-time is causal and the stability of the same is ensured in a suitable manner. A stably causal M is strongly causal and hence causality determines the topology of the space-time. Finally, a stably causal space-time admits the existence of a *global time function* on M, that is, a smooth scalar field f, for which the gradient is always timelike (Hawking and Ellis, 1973). Such a function assigns a value of 'time' to each point in M so that this time is strictly increasing along all timelike curves. This is precisely the property one desires for a global assignment of a time coordinate in M. Of course, such a time function is not unique. In Section 4.6 we will discuss one such function which increases along all timelike curves when M is stably causal.

4.3 Reflecting space-times

The space-time of Fig. 26, through stably causal, has many undesirable features which may not be regarded as physical. The point q is causally connected to events arbitrary close to p but p receives no signal from q or the region $I^+(q)$. Next, even a small perturbation in the metric results in a sudden increase in $I^+(p)$, or a small displacement from p results in an abrupt increase in the volume of future of an event. Thus, even though the space-time could be stably causal, one may want to impose further conditions on M so as to exclude the above features.

Such a requirement was proposed in the form of reflectingness of a space-time by Hawking and Sachs (1974). A space-time (M, g) is called reflecting if, for all p and q in M,

$$I^+(p) \supseteq I^+(q) <=> I^-(p) \subseteq I^-(q). \tag{4.17}$$

The reflectingness of M can be characterized in terms of the chronological common futures and pasts defined earlier, as shown by Hawking and Sachs (1974): (M, g) is reflecting if and only if $\uparrow I^-(p) = I^+(p)$ and $\downarrow I^+(p) = I^-(p)$ for all $p \in M$. Clearly, the relations

$$I^+(p) \subseteq \uparrow I^-(p),$$

$$I^-(p) \subseteq \downarrow I^+(p),$$

are true in general whether M is reflecting or not. Note that $\uparrow I^-(p)$ and $\downarrow I^+(p)$ are always open sets in M just as $I^-(p)$ and $I^+(p)$ are.

In the case of space-time of Fig. 26, $I^-(p) \subseteq I^-(q)$ but $I^+(p) \supseteq I^+(q)$ is not true. In fact, we have $I^+(q) \cap I^+(p) = \emptyset$. This situation, in which p lies on the boundary of $I^-(q)$, is decisive as to whether or not a space-time is reflecting. This is made precise by the following result.

Proposition 4.16. The following are equivalent:

(a) (M, g) is reflecting

(b) $q \in \dot{I}^+(p)$ iff $p \in \dot{I}^-(q)$.

Proof. To show that (b) implies (a), suppose $I^+(p) \supseteq I^+(q)$. Then if $p \ll q$ the situation is trivial, so choose them not chronologically related, giving

$q \in \dot{I}^+(p)$ and hence $p \in \dot{I}^-(q)$, which, by definition of $\dot{I}^-(q)$, implies that $I^-(p) \subseteq I^-(q)$. Similarly, $I^-(q) \supseteq I^-(p)$ implies $I^+(q) \subseteq I^+(p)$.

To see the converse, let $q \in \dot{I}^+(p)$ and suppose M to be reflecting. Then $I^+(q) \subseteq I^+(p)$ and hence $I^-(p) \subseteq I^-(q)$. Since p is not chronologically-related to q, this gives $p \in \dot{I}^-(q)$. The other part can also be shown, using the reflecting condition in a similar way. □

The reflecting condition on M is related in an interesting manner with the causality of M. It implies that either chronology is fully respected or it is violated as badly as possible as shown by the following (Clarke and Joshi, 1988).

Proposition 4.17. Let (M, g) be a reflecting space-time containing a closed timelike curve through $p \in M$. Then there is a closed timelike curve through every point of M.

Proof. We show that $I^+(p) = I^-(p) = M$, which implies that $x \ll p \ll x$ for all points x. Suppose, on the contrary, that $I^+(p) \neq M$ and take $q \in \dot{I}^+(p)$. Then by Proposition 4.16, $p \in \dot{I}^-(q)$. Since $p \ll p$, $I^+(p)$ is an open neighbourhood of p and so there is a point $s \in I^+(p) \cap I^-(q)$, that is, $p \ll s \ll q$. But then $q \in I^+(p)$, contradicting the assumption. So $I^+(p) = M$. The proof that $I^-(p) = M$ is exactly dual to this, so the result follows. □

In view of this, the reflectingness of M rules out non-universal causality violation such as that occurring in the Kerr metric with $a < m$ below the horizon. However, global causality violation (for example, Kerr metric with $a > m$) is allowed. Is there a physically interesting class of space-times which is reflecting? The following (Clarke and Joshi, 1988) provides an answer.

Proposition 4.18. Any stationary space-time is reflecting.

Proof. Let ϕ_t be a family of diffeomorphisms corresponding to a future directed timelike Killing vector. Suppose $I^+(q) \subseteq I^+(p)$. Then for all $t > 0$ we have $\phi_t(q) \in I^+(q)$ and so $p \in I^-(\phi_t(q))$. Applying ϕ_{-t} to this and noting that ϕ_{-t} preserves causal relations gives the result $\phi_{-t}(p) \in I^-(q)$. Now, let $x \in I^-(p)$. Then for small enough $\delta > 0$ we have

$$x \in I^-(\phi_{-\delta}(p)) \subseteq I^-(q),$$

from the above. Thus $I^-(p) \subseteq I^-(q)$. The proof of the reverse implication is the exact dual, so (M, g) is reflecting. □

It was indicated by Hawking and Sachs (1974) that basically the violation of reflectingness of a space-time should result from cutting out closed sets from M of 'dimension' more than two. The point is, one could always start with a regular space-time, cut out any closed set and make suitable identifications. One still has a manifold with a Lorentzian metric. This would give a space-time which is incomplete in a certain sense. The portions one could cut out could be points or higher dimensional sets from M. It is not difficult to see that if a point is cut out from Minkowski space-time, the result is still a reflecting space-time. This crucial property is classified as below. Let S be a closed subset of M with Hausdorff dimension less than three and let (M, g) be reflecting. Consider now the space-time $(M', g \mid M')$ where $M' = M - S$. Then the following can be shown.

Lemma 4.19 (Clarke and Joshi, 1988). If $x, y \in M'$ then $x \in I^-(y, M)$ if and only if $x \in I^-(y, M')$.

The following now characterizes the higher dimensional cuts in a space-time.

Theorem 4.20. If S is a closed subset of M with Hausdorff dimension less than three and (M, g) is reflecting, then $(M', g \mid M')$ is also reflecting with $M' = M - S$.

Proof. Suppose $p, q \in M'$ satisfy $I^-(p, M') \subseteq I^-(q, M')$. If $x \in I^-(p, M) \cap M'$ then Lemma 4.19 implies $x \in I^-(p, M')$. Then, this implies, by assumption that $x \in I^-(q, M')$. Then Lemma 4.19 again implies $x \in I^-(q, M)$. That is, $I^-(p, M) \cap M' \subseteq I^-(q, M)$. Now, using the fact that $I^-(p, M)$ is open and M' is dense we get,

$$I^-(p, M) = Int\,\overline{I^-(p, M)} = Int\,\overline{I^-(p, M) \cap M'}.$$

But this later set is contained in $Int\,\overline{I^-(q, M)} = I^-(q, M)$. So, since M is reflecting, $I^+(p, M) \supseteq I^+(q, M)$. A similar argument then shows that this implies

$$I^+(p, M') \supseteq I^+(q, M').$$

The proof of the converse implication is identical. This implies that $(M', g \mid M')$ is reflecting. $\square$

The removal of a three-plane can, however, cause a violation of reflecting condition if the plane is spacelike, as shown by the space-time of Fig. 26. It is relevant to ask here if a violation of the reflecting condition is always associated with a 'missing' three-surface. That this is not necessarily so is shown by an example given by Clarke and Joshi (1988) where a removal of a two-dimensional set from Minkowski space-time gives rise to a non-reflecting space-time. Basically, if the reflecting condition is violated, then $I^+(p)$ has a 'shadow' cast in it by the singularity structure. The following (where the set O is the shadow) makes this precise.

Proposition 4.21. Let (M, g) be a non-reflecting space-time with $I^+(p) \not\supset I^+(q)$, but $I^-(p) \subset I^-(q)$. Then the set $O = Int(\uparrow I^-(p) - I^+(p))$ is non-empty.

Proof. From Proposition 4.16 (or directly) we have the implication of violation of reflecting condition that the set $A = \uparrow I^-(p) - I^+(p)$ is non-empty. Suppose now that $O = IntA$ is empty. Then for any $q \in A$ and any neighbourhood N of q we must have $N_q \cap I^+(p) \neq \emptyset$ and so $q \in \dot{I}^+(p)$. This implies $\uparrow I^-(p) \subseteq \overline{I^+(p)}$. Taking interiors,

$$\uparrow I^-(p) \subseteq Int\overline{I^+(p)} = I^+(p). \tag{4.18}$$

But we always have $I^+(p) \subseteq \uparrow I^-(p)$ which gives from the above, $I^+(p) = Int \uparrow I^-(p)$, contradicting the fact that A is non-empty. □

We note that a reflecting space-time need not be causal; for example, the manifold $M = S^1 \times \mathbb{R}$ discussed in Section 4.2 is reflecting but not causal. The reflecting condition basically characterizes the existence or otherwise of higher dimensional 'cuts' in a space-time, ensuring the well-behaved nature of M in that respect as discussed earlier in this section. However, in order to ensure the causal regularity of M, an additional causality assumption is needed. Hawking and Sachs (1974) define a causally continuous space-time as a pair (M, g) when M is reflecting and also satisfies the distinguishingness condition. They showed that causal continuity of M removes many global pathologies and introduces desirable features like continuity of global time functions and outer continuity of the maps I^+ and I^- in M. Further, it was shown by Budic and Sachs (1974) that when M is causally continuous, the causality structure of space-time extends to its ideal points boundary as well, namely the points at infinity and singularities.

It can be seen (Akolia, Joshi and Vyas, 1981) that when M is causally continuous, one does not require different causality conditions as given in Section 4.2 in the sense that they become equivalent to each other (except causality), and one has a unified statement of causality principle for M. We recall that M is called future distinguishing if $I^+(x) = I^+(y)$ implies $x = y$ for all x, y in M and a past distinguishing space-time is similarly defined. In general M can be future distinguishing but may violate past distinguishingness and vice versa. The space-time is distinguishing if it is both future and past distinguishing. Even though causal continuity assumes distinguishingness, actually either the future or past condition will be sufficient. Because, when M is reflecting, it is easy to see using the definition that $I^+(x) = I^+(y)$ if and only if $I^-(x) = I^-(y)$ for all points x, y in M. Thus, we have the following.

Proposition 4.22. A space-time is causally continuous if and only if it is reflecting and either future or past distinguishing.

The structure of Seifert future $J_S^+(x)$ turns out to be particularly simple for a causally continuous space-time as shown by the following result.

Proposition 4.23 (Hawking and Sachs 1974). Let M be a causally continuous space-time. Then, for any x in M

$$J_S^+(x) = \overline{J^+(x)}.$$

Again, by Proposition 4.13, $Int[A^+(x)] = \uparrow I^-(x)$ in general. For a reflecting space-time, taking closure on both sides gives

$$A^+(x) = \overline{I^+(x)}. \tag{4.19}$$

Since $\overline{I^+(x)} = \overline{J^+(x)}$, for a reflecting space-time we get

$$A^+(x) = J_S^+(x). \tag{4.20}$$

We have shown in Section 4.2 that stable causality implies W-causality of M in general. Now the above expression shows that a W-causal reflecting space-time must be stably causal, and hence W-causality and stable causality are equivalent conditions for such a space-time. Next, suppose $I^+(x) = I^+(y)$. Then by the definition of almost causal precedence, we have both xAy and yAx. Again, when xAy and yAx are both true, eqn (4.19) gives $y \in \overline{I^+(x)}$ and $x \in \overline{I^+(y)}$. This implies $I^+(y) \subseteq I^+(x)$ and $I^+(x) \subseteq I^+(y)$, that is, $I^+(x) = I^+(y)$. It follows that when M is reflecting, $I^+(x) = I^+(y)$ if and only if xAy and yAx. Suppose now M is causally continuous and xAy and yAx holds. Then by the above, $I^+(x) = I^+(y)$. Then distinguishingness implies $x = y$ which shows that M is W-causal. Thus, we have proved the following.

Theorem 4.24. Let (M, g) be a reflecting space-time. Then the following are equivalent:

(a) M is future distinguishing,
(b) M is past distinguishing,
(c) M is distinguishing,
(d) M is W-causal,
(e) (M, g) is stably causal,
(f) M is strongly causal.

Thus, for a reflecting space-time only the future or past distinguishingness becomes the single reasonable global causality condition with no need to assume any further causal requirement. This is not the case for a general space-time where there is a separate need for each causality condition.

We note that the causality or chronology condition is not included in the list above in Theorem 4.24. In fact, causality is strictly weaker than any of the above conditions even for a reflecting space-time. Consider, for example, the space-time of Fig. 24. Here M is reflecting and causal but not distinguishing. We recall that reflecting condition actually rules out only higher dimensional cuts from M. For example, if we remove a point from the Minkowski space-time, or for that matter infinitely many discrete isolated points, M will still be reflecting. However, as shown in Fig. 27, such a situation may not be called physical in the following sense. If a point is removed, $J^+(p)$ may not be closed in M and one has the situation that p causally precedes events q_n for all n but not the event q which is the limit point of the sequence q_n. Thus p sends messages to events arbitrarily close to q but no message exists between p and q. To avoid such a situation, we say that M is *causally simple* if the causal futures and pasts $J^{\pm}(x)$ are closed for all x in M. We note that causality of M is not included in the definition of causal simplicity here. Physically interesting space-times such as the Schwarzschild, Kruskal–Szekeres extension, Robertson–Walker models and so on, satisfy this condition. One can see using the definition of reflectingness that a causally simple space-time must always be reflecting and thus causal simplicity is a strictly stronger condition between the two. It can be shown now that for a causally simple space-time the causality condition is characterized in terms of boundaries of the future and past of an event x.

Proposition 4.25. Let the causal futures (pasts) $J^{\pm}(x)$ be closed in M. Then M is causal if and only if $\dot{I}^+(x) \cap \dot{I}^-(x) = \{x\}$ for all x in M.

Proof. let M be causal and suppose $y \in \dot{I}^+(x) \cap \dot{I}^-(x)$ such that $y \neq x$. Since $\dot{I}^+(x) = \dot{J}^+(x)$ is true in general, $y \in \dot{J}^+(x)$ and so $y \in J^+(x)$ since $J^+(x)$ is a closed set in the manifold topology. Thus there is a future directed causal curve from x to y. In the same way, $y \in \dot{I}^-(x)$ implies a future directed causal curve from y to x. This yields a closed causal curve through x, violating causality. Thus $x = y$.

Conversely, if $\dot{I}^+(x) \cap \dot{I}^-(x) = \{x\}$, then $x \in \dot{I}^+(x)$. Now $I^+(x)$ is always open and so $x \notin I^+(x)$ and similarly $x \notin I^-(x)$. This avoids closed timelike curves. Again, if there were a closed non-degenerate null geodesic through x then for some $y \neq x$, and y on this null geodesic, we have both $y \in \dot{I}^+(x)$ and $y \in \dot{I}^-(x)$, which is not allowed. Hence M must be causal. □

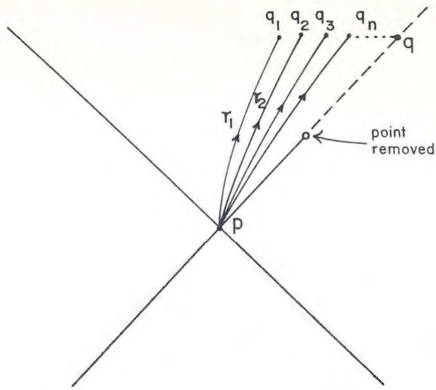

Fig. 27 The event p is causally connected to events arbitrarily near to q, but there is no causal connection between p and q.

Next, under the given assumptions it is now possible to see that the above boundary condition is equivalent to the distinguishingness of the space-time. For that, if M is distinguishing then it is causal and by Proposition 4.25 the topological boundary condition holds. Conversely, let $\dot{I}^+(x) \cap \dot{I}^-(x) = \{x\}$ be satisfied for all $x \in M$ and suppose $I^+(x) = I^+(y)$. Then $\dot{I}^+(x) = \dot{I}^+(y)$ and thus $x \in \dot{I}^+(y)$. But we also have $\dot{I}^+(y) \cap \dot{I}^-(y) = \{y\}$, which gives $y \in \dot{I}^+(x)$. Now $\dot{I}^+(x) = \dot{J}^+(x)$ and the closure of $J^+(x)$ implies $x \in J^+(y)$ and $y \in J^+(x)$, which is a violation of causality, hence a contradiction with Proposition 4.25. Thus we must have $x = y$. It can be similarly shown that $I^-(x) = I^-(y)$ also implies $x = y$. Thus we have shown that the topological boundary condition given above characterizes the distinguishingness of the space-time.

Suppose now M is causally simple and the topological boundary condition above is satisfied. Then M is reflecting and it is causal and distinguishing as seen above. However, in such a case, all the higher order causality conditions are equivalent as shown by Theorem 4.24, and these in turn will be equivalent to the causality of space-time as shown above. Thus, when M avoids global pathological features as characterized here, the topological boundary condition above provides a single unified characterization for the causal regularity of M, and separate causality conditions are not needed.

4.4 Global hyperbolicity

A space-time (M, g) is said to be *globally hyperbolic* if the sets $J^+(x) \cap J^-(y)$ are compact for all x, y in M and M is causal. It is not difficult to see that a globally hyperbolic space-time must be causally simple.

Proposition 4.26. Let (M,g) be globally hyperbolic. Then the sets $J^{\pm}(x)$ are closed for all $x \in M$.

Proof. Suppose $J^+(x)$ is not closed and let $y \in \overline{J^+(x)}$ but $y \notin J^+(x)$. Let $z \in I^+_{\cdot}(y)$. Then $y \in \overline{J^+(x) \cap J^-(z)}$ but $y \notin J^+(x) \cap J^-(z)$. This is a contradiction because $J^+(x) \cap J^-(z)$ is closed, being compact. □

It is possible to generalize the above to show that if S is any compact set, $J^+(S)$ will be closed in M and if S_1 and S_2 are any two compact subsets, $J^+(S_1) \cap J^-(S_2)$ must be compact in M. From the results of the last section it follows now that if M is globally hyperbolic then it is strongly and stably causal, and is reflecting and causally continuous. In fact, as we will see, global hyperbolicity is a rather strong condition on M which uniquely fixes the overall topology of the space-time. The space-time then has a very regular global behaviour and all the pathological features stated in last section are removed.

Physically interesting space-times such as the Schwarzschild solution, Friedmann–Robertson–Walker cosmological solutions, and the steady state models are all globally hyperbolic. It is thus useful to study this general class which includes these models. We review several properties of these space-times here and also study the domain of dependence for a closed achronal set S and the associated Cauchy horizon which develops when M is not globally hyperbolic. As we will see in Chapter 6, another strong motivation for global hyperbolicity is the cosmic censorship hypothesis of Penrose which rules out timelike (naked) singularities. It turns out that if one rules out even the locally naked singularities, the resulting space-time structure must be globally hyperbolic. In the absence of cosmic censorship, signals from space-time singularities could play havoc with the deterministic structure of space-time.

The global hyperbolicity of M is closely related to the future or past development of initial data from a given spacelike hypersurface. First, we give here some definitions. Let S be a closed achronal set. The *edge* of S is defined as a set of points $x \in S$ such that every neighbourhood of x contains $y \in I^+(x)$ and $z \in I^-(x)$ with a timelike curve from z to y which does not meet S. When S is a closed achronal set without an edge one has a result similar to Proposition 4.3 that S is a three-dimensional, embedded, C^0 submanifold of M. A *partial Cauchy surface* S is defined as an acausal set without an edge. Thus, no non-spacelike curve intersects S more than once and S is a spacelike hypersurface. The *future domain of dependence* of S, denoted by $D^+(S)$, is defined as the set of all points $x \in M$ such that every past inextendible non-spacelike curve from x intersects S. It is clear that $S \subset D^+(S) \subset J^+(S)$ and S being achronal, $D^+(S) \cap I^-(S) = \emptyset$. We note that Penrose (1972) and Geroch (1970a) use timelike curves to

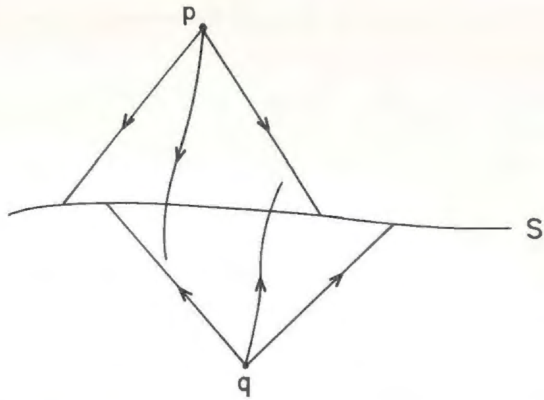

Fig. 28 The spacelike hypersurface S is a Cauchy surface in the sense that for any p in future of S, all past directed non-spacelike curves from p intersect S. The same holds for all future directed curves from any point q in past of S.

define the domain of dependence, rather than non-spacelike curves used above, which agrees with Hawking and Ellis (1973). The *past domain of dependence* $D^-(S)$ for a partial Cauchy surface S is defined dually. The full *domain of dependence* for S is defined as

$$D(S) = D^+(S) \cup D^-(S). \tag{4.21}$$

A partial Cauchy surface is called a *Cauchy surface* or a *global Cauchy surface* if $D(S) = M$. Clearly, for a Cauchy surface S, $edge(S) = \emptyset$. Every non-spacelike curve in M must meet S once (and exactly once) if S is a Cauchy surface. The relationship between the global hyperbolicity of M and the notion of Cauchy surface is the following (we refer to Hawking and Ellis, 1973, for the proof).

Theorem 4.27. A space-time M is globally hyperbolic if and only if it admits a spacelike hypersurface S which is a Cauchy surface for M (Fig. 28).

We note that even when M is globally hyperbolic, all spacelike surfaces in M need not be Cauchy surfaces. For example, for the Minkowski space-time, the spacelike surfaces $t =$ const. are global Cauchy surfaces, but the hyperboloids

$$t^2 - x^2 - y^2 - z^2 = \text{const.},$$

are not because the past or future null cones of the origin are boundaries of domain of dependence for these spacelike surfaces. Again, as shown by Fig.

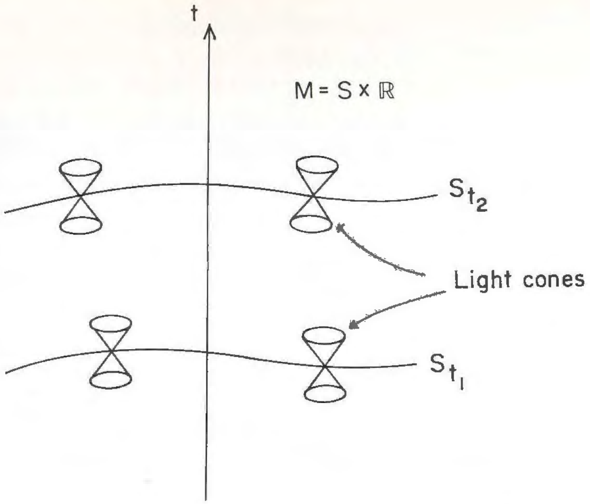

Fig. 29 M is globally hyperbolic with topology $M = S \times \mathbb{R}$. For any $t \in \mathbb{R}$, the spacelike hypersurface S_t of constant time is a Cauchy surface for the space-time.

24, if one deletes a point from the Minkowski space-time, the resulting M admits no Cauchy surface and M is not globally hyperbolic. The following theorem, clarifying the structure of a globally hyperbolic space-time, was given by Geroch (1970a) and shows that a globally hyperbolic M has a unique topological structure and it admits no topology change.

Theorem 4.28. Let M be globally hyperbolic, then M is homeomorphic to $\mathbb{R} \times S$ where S is a three-dimensional submanifold and for each $t \in \mathbb{R}$, $\{t\} \times S$ is a Cauchy surface for M (Fig. 29).

The basic idea of the proof for the above involves introducing a finite measure μ on M so that $\mu(M) = 1$. Then a function $f : M \to \mathbb{R}$ is introduced by

$$f(p) = \frac{\mu(J^+(p))}{\mu(J^-(p))}.$$

The sets $f =$ const. are seen to be Cauchy surfaces for M. This is closely related to the idea of causal functions on a space-time, which we discuss in detail in Section 4.6.

One would like to ask if the converse of Theorem 4.27 is true in some sense, that is, whether the direct product space-times (M, g) with $M = S \times T$, where each $S \times \{t\}$ is spacelike and each $\{x\} \times T$ is timelike, are always globally hyperbolic. The answer is in negative and we refer to Clarke and Joshi (1988) for an example of a space-time which is a direct product as above but not globally hyperbolic, and in fact, which is not even reflecting.

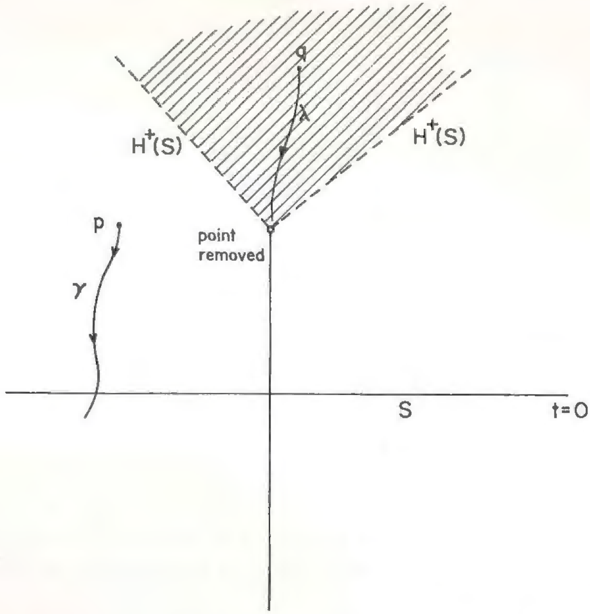

Fig. 30 The space-time obtained by removing a point from the Minkowski space is not globally hyperbolic. The point q does not belong to $D^+(S)$ as there are non-spacelike curves like λ which do not meet S in the past. The event $p \in D^+(S)$. The Cauchy horizon is the boundary of the shaded region which consists of points not in $D^+(S)$.

Next, let S be a partial Cauchy surface. Then $N = D^+(S) \cup D^-(S) \neq M$ and N must be a proper subset of M. The boundary of N in M can be divided into two regions $H^+(S)$ and $H^-(S)$ which are respectively called the *future* and *past Cauchy horizons* of S. We can write

$$H^+(S) = \{x \mid x \in D^+(S), I^+(x) \cap D^+(S) = \emptyset\}.$$

The past horizon $H^-(S)$ is defined in a dual manner. Even though M may not be globally hyperbolic and S is not a Cauchy surface, the region $Int(D^+(S))$ or $Int(D^-(S))$ is globally hyperbolic in its own right and the surface S serves as a Cauchy surface for the manifold $Int(N)$. Thus, $H^+(S)$ or $H^-(S)$ represent the failure of S to be a global Cauchy surface for M (see Fig. 30). The following result illustrates the nature of $H^+(S)$.

Proposition 4.29. $H^+(S)$ is achronal and closed.

Proof. Suppose $x, y \in H^+(S)$ with $x \ll y$. Then $x \in I^-(y)$ and there exists a neighbourhood $N_x \subset I^-(y)$. Let $p \in N_x \cap I^+(x)$. Then $p \ll y$

and every past directed timelike curve from p can be extended to be a past directed curve from y which must meet S. Thus $p \in D^+(S)$, which is a contradiction to $D^+(S) \cap I^+(x) = \emptyset$. Hence, no two points of $H^+(S)$ are timelike related. Next, let x be a limit point of a sequence of points $\{x_n\}$ in $H^+(S)$. Suppose γ is a past directed timelike curve not meeting S, with future end point at x. Then, for any $y \in \gamma$, $I^+(y)$ contains all points $\{x_n\}$ for some $n \geq k$. Then coming from x_n to y and following γ gives a non-spacelike curve from x_n not meeting S, which is contradictory to $x_n \in D^+(S)$. $\square$

As in the case of the boundary of a future set, $H^+(S)$ is generated by null geodesics which are either past inextendible in $H^+(S)$ or have a past end point on the edge of S (Geroch, 1970a).

The basic physical relevance of global hyperbolicity condition on M is that it implies a deterministic structure for the space-time in the sense characterized by the above results, given an initial data defined on a Cauchy surface S. In fact, the notion of global hyperbolicity was introduced by Leray (1952) and it is seen from considerations on the Cauchy problem in general relativity (see for example, Wald, 1984 for a review) that if N is a globally hyperbolic subset of M, the wave equation for a δ-function source at any $p \in N$ has a unique solution which vanishes outside $N - J^+(p)$. The definition of global hyperbolicity given by Leray involves the space $C(p,q)$, which is the set of all C^0 non-spacelike curves from p to q which are same up to a reparametrization. Defining a C^0 topology on this space of curves by stating that two curves are nearby if their points in M are close enough allows the global hyperbolicity to be characterized in term of $C(p,q)$ as seen from the following proposition.

Proposition 4.30 (Seifert 1967). Let M be a strongly causal space-time. Then M is globally hyperbolic if and only if $C(p,q)$ is compact for all $p, q \in M$.

The set of all C^1 timelike curves from p to q forms a dense subset of $C(p,q)$. The length of any such curve is defined by eqn (2.33) where t is a C^1 parameter which could be chosen to be the proper time along the curve.

An important property of globally hyperbolic space-times, or any globally hyperbolic subset in M, which is relevant for the singularity theorems is the existence of maximum length non-spacelike geodesics between pairs of causally related events. In a complete Riemannian manifold with a positive definite metric, any two points can be joined by a geodesic of minimum length and in fact such a geodesic need not be unique. The analogue of this result for Lorentzian metrics was given by Avez (1963) and Seifert (1967):

Theorem 4.31. Let (M, g) be globally hyperbolic and $p, q \in M$ such that $p < q$. Then there is a non-spacelike geodesic from p to q whose length

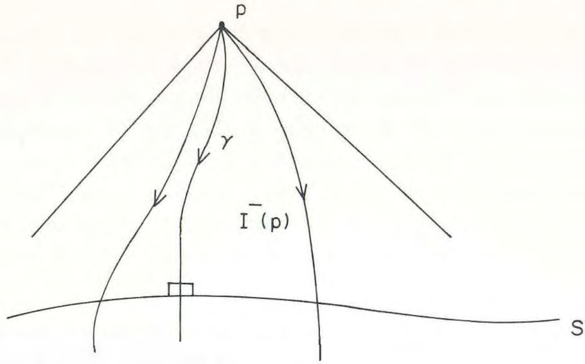

Fig. 31 In a globally hyperbolic space-time, given a Cauchy surface S and a point p, there exist timelike geodesics such as γ orthogonal to S, which maximize the lengths of all non-spacelike curves from p to S.

is greater than or equal to that of any other future-directed non-spacelike curve from p to q.

Thus, in globally hyperbolic space-times, there is a finite upper bound on the proper time lengths of non-spacelike curves between two chronologically related events. It is clear that there is no lower limit of lengths for such curves except zero, because the chronologically related events can always be joined using broken null curves which could give an arbitrary small length curve between them. Similarly, if S is a Cauchy surface in a globally hyperbolic space-time M, then for any point p in the future of S, there is a past directed timelike geodesic from p orthogonal to S which maximizes the lengths of all non-spacelike curves from p to S (Fig. 31).

We address here finally the following interesting question. The minimal length result stated above for positive definite metrics holds for all complete Riemannian manifolds. The corresponding Lorentzian result given in Theorem 4.31 is valid, however, only for globally hyperbolic space-times. Now, given two chronologically related events, an observer can commute between them along a timelike curve. The important question here is how much proper time the observer will spend along the timelike curve before the second event is reached. It would appear physically reasonable to demand that any timelike signal from p should not take more than a certain finite and bounded amount of proper time before reaching q. This is the property realized for globally hyperbolic space-times as shown by Theorem 4.31. However, global hyperbolicity is a rather strong condition on M (see for example, Hawking and Sachs, 1974) and it would be useful to know if the physically appealing property of having a finite upper bound for the lengths of non-spacelike curves still holds when M is not globally hyperbolic but satisfies a weaker causality condition.

In the following, we show by means of a series of examples that, given two chronologically related events, no other causality condition weaker than global hyperbolicity is sufficient to ensure an upper bound to the proper time lengths of timelike curves between them (Dhurandhar and Joshi, 1981). Specifically, we give here following examples of space-times: (a) a chronological M that contains a pair of events p, q such that $p \ll q$ and the finite upper bound property is violated for timelike curves from p to q, (b) a causal space-time with a similar property, and (c) a causally simple M with this property. Since a causally simple and causal space-time is strongly and stably causal, causally continuous, and obeys Carter's causality conditions, the conclusion stated above follows.

Suppose first that chronology is violated at an event p. Then it is possible to come back to p along a future directed timelike curve, say, γ. If one considers some other point $q \in I^+(p)$, then it is possible to traverse along γ an arbitrary number of times and then reach q by some causal curve. Now γ, being a timelike curve, has non-zero length, and so one gets a future directed timelike curve from p to q whose length exceeds any preassigned positive number. Hence $d(p, q)$ is infinite, where $d(p, q)$ denotes the maximum of lengths for all non-spacelike curves from p to q.

Next, consider the situation when chronology holds but causality fails in a space-time. Consider the two-dimensional Minkowski space-time with metric $ds^2 = -dt^2 + dx^2$ and use the new coordinates

$$u = t - x, \qquad v = t + x,$$

and restrict ourselves to the region $u \geq 0, v \geq 0$. The light cones are given by $u = const, v =$ const. values. We make the following identification in the space-time. For all $u_0 \in [0, 1)$ and all v, $u_0 \sim u_0 + n$, with $n = 1, 2, ...$, that is, all positive integers. This identification can also be achieved by defining a new coordinate with a range $-\pi < \theta < \pi$, by defining $\theta = 2 \sin^{-1}(2u - 1)$. This defines our space-time. The series of curves $u = (n - \frac{1}{2}v)$, $n = 1, 2...$, from the point $p = (0, 0)$ to $q_n = (n - \frac{1}{2}, 1)$ in the unidentified space have lengths which grow without bound as $n \to \infty$. The nth curve here has the length $(n - \frac{1}{2})^{1/2}$. Following the identification the points q are identified with the point, say, $q = q_1$, and since lengths do not change after identification, the lengths of curves from p to q have no upper bound as $n \to \infty$. The situation is as shown in Dhurandhar and Joshi (1981, Fig. 1a and 1b). There is a closed null geodesic through p due to the identification and the space-time is chronological but not causal.

To consider the case when causality holds but not the distinguishingness, the space-time of Fig. 24 can be used where this happens. For our purpose we define the metric appropriately. For the sake of definiteness, the space-time is given by $0 \le t \le 1$, $0 \le x \le 1$ and the lines $x = 0$ and $x = 1$ are identified and the point $x = \frac{1}{2}, t = 0$ is removed. The lower light cone arm has zero slope on the $t = 0$ axis and slowly tilts up as t increases. If we take $p \equiv (0,0)$ and $q \equiv (1/2, 1)$ then p can be joined to q by non-spacelike curves which can make an arbitrary number of 'rounds' by making the slope of such a curve sufficiently small. But this does not guarantee that the arc length of such curves would grow without bound. The tangent to such a curve approaches the null direction making the contribution to the total length from each 'lap' very small. In other words, there is a competition between the number of laps and the length of each lap. However, one can get the desired result of unbounded length by choosing the metric appropriately. The following choice of the metric function is satisfactory:

$$ds^2 = (2dx - dt)(dt - t^2 dx). \tag{4.22}$$

This defines the light cone field on the space-time with the upper arm having slope 2 and the lower arm having slope t^2. On the x-axis the lower arm touches the axis. For our calculations we consider the strip $0 \le x < \infty$, $0 \le t \le 1$ and then make the necessary identification to retrieve the above example. The sequence of curves we take are geodesics from $p \equiv (0,0)$ to each of the points $(n + \frac{1}{2}, 1), n = 1, 2, ...$ We show that these non-spacelike curves have the required property.

The first integrals of the geodesics are given by

$$(t^2 + 2)\frac{dt}{ds} - 4t^2\frac{dx}{ds} = k_n,$$

$$\left(2\frac{dx}{ds} - \frac{dt}{ds}\right)\left(\frac{dt}{ds} - t^2\frac{dx}{ds}\right) = 1,$$

where k is an arbitrary constant and is determined by the fact that the geodesic joins $(0,0)$ to $(n + \frac{1}{2}, 1)$. From the expression for dx/dt obtained from the above, we have

$$n + \frac{1}{2} = \int_0^1 \frac{1}{4t^2}\left[2 + t^2 - \frac{k_n(2-t^2)}{(k_n^2 + 8t^2)}^{1/2}\right] dt.$$

This relation determines k_n. This integral can be written as

$$\frac{1}{4} + \frac{k_n}{8\sqrt{2}}\log\left[\frac{\sqrt{8}}{k_n} + \left(1 + \frac{8}{k_n^2}\right)^{1/2}\right] + \frac{1}{2}\int_0^1 \frac{1}{t^2}\left[1 - \left(1 + \frac{8t^2}{k_n^2}\right)^{-1/2}\right] dt.$$

Therefore as $k_n \to 0$, the integral diverges as $1/k$. So,

$$k_n \sim \frac{1}{\sqrt{n}} \quad as \quad n \to \infty.$$

The total arc length of the nth curve is given by

$$s_n = \int_0^1 \left[\frac{2-t^2}{(k_n^2+8t^2)}\right]^{1/2} dt \sim log\frac{1}{k_n} \quad as \quad k_n \to 0.$$

From these two results we have

$$s_n \sim \log n \quad \text{as} \quad n \to \infty. \tag{4.23}$$

The identification as earlier and the removal of the point $(1/2, 1)$ ensures that the space-time is causal but not distinguishing. The lengths of the curves do not alter after the identification so this family of curves which join the point $p \equiv (0,0)$ and $q \equiv (1/2, 1)$ in the new space-time have arc lengths which grow without bound. Hence $d(p,q)$ is infinite for this choice of p and q. It should be noted that if we had taken the slope of the lower light-cone arm to be simply t, then the arc lengths of the geodesics would have been bounded, as one can easily see by going through a similar calculation. It is important that the lower light-cone arm tilts with sufficient slowness to get the required result.

We now consider a causally simple space-time which is not globally hyperbolic. The space-time given by eqn (4.22) may be used to construct a model which has this property and in which the lengths of causal curves from p to q would grow without bound. We alter only the identification. For all $t_0 \in [\frac{1}{2}, 1]$ we identify the point (x_0, t_0) with the points $(x_0 + n, t_0), n = 1, 2, \ldots.$ It is seen that the sets $J^{\pm}(p)$ are closed for each event p but $J^+(p) \cap J^-(q)$ is not compact for, say, $p = (0,0)$ and $q = (\frac{1}{2}, 1)$. So the space-time is causally simple but not globally hyperbolic. Again, here $d(p,q)$ is infinite.

Let M be globally hyperbolic and $p \ll q$, then as stated, there is at least one timelike geodesic of finite length from p to q along which the length is maximized. It is necessary to check if global hyperbolicity is too strong a condition and if a weaker causality condition can replace it while preserving the above property. We have shown above that this is indeed not the case and there is no finite upper bound on lengths of non-spacelike curves when global hyperbolicity is replaced by a weaker causality condition. Now global hyperbolicity implies the 'upperbound property', which in turn would imply the chronology of the space-time. However, this upperbound property does not imply causality. The space-time given by the metric

$$ds^2 = (2dx - dt)(dt - tdx), \tag{4.24}$$

where $0 \leq x \leq 1$ and $0 \leq t \leq 1$ with $t = 0$ and $t = 1$ identified, is not causal but has the upper-bound property.

4.5 Chronal isomorphisms

When we compare two models of the universe, the space-time isomorphisms play an important role in carrying properties of one model into the other one. Let (M, g) and (M', g') be two space-times. A one-one and onto map ϕ from M to M' is a *chronal isomorphism* if for all $x, y \in M$,

$$x \ll y \iff \phi(x) \ll \phi(y). \tag{4.25}$$

Thus, ϕ is a bijection which preserves the chronology relation in both directions. The definition given above is slightly different from Budic and Sachs (1974), who assume $x < y$ if and only if $\phi(x) < \phi(y)$. When $M = M'$, then this is the same as the automorphisms introduced by Zeeman (1964) and Lester (1984). The difference is that, these later authors deal either with the Minkowski or deSitter case whereas we consider here a general space-time. Such maps have been studied by Vyas and Akolia (1984) and Vyas and Joshi (1989) and this definition is the same as that of the causal isomorphism of Malament (1977), but since it preserves chronal precedence we call it a chronal isomorphism.

Chronal automorphisms are of physical interest because, as shown by Alexandrov (1953, 1967) and Zeeman (1967), the bijections of a Minkowski space-time which preserve a null separation in both directions must be Lorentz transformations up to translations and dilatations. Thus, the group of chronal automorphisms is generated by the inhomogeneous Lorentz group and dilatations. The following result shows that the map ϕ and the operations I^+ (or I^-) commute.

Proposition 4.32. Let $\phi : (M, g) \to (M', g')$ be a chronal isomorphism. Then $\phi(I^+(x)) = I^+(\phi(x))$ for all x in M and a similar result holds for I^-.

Proof. Let $w \in \phi(I^+(x))$. Then there is $z \in I^+(x)$ such that $\phi(z) = w$ But $z \gg x$ and $\phi(z) = w$ implies from the definition of chronal isomorphism that $w \gg \phi(x)$, or $w \in I^+(\phi(x))$. Thus, we have $\phi(I^+(x)) \subseteq I^+(\phi(x))$. Conversely, let $s \in I^+(\phi(x))$. Then $s \gg \phi(x)$ and there is $t \in M$ such that $\phi(t) = s$ and $\phi(t) \gg \phi(x)$ because ϕ is onto. Then $t \gg x$ by the definition of chronal isomorphism which again implies $\phi(t) \in \phi(I^+(x))$, $\phi(t) = s$. So we have $s \in \phi(I^+(x))$ and thus $I^+(\phi(x)) \subseteq \phi(I^+(x))$, which proves the required equality. □

As noted in Section 4.2, the topological structure of M is determined by causal relations in the form of the Alexandrov topology on M, generated by open sets of the form $I^+(x) \cap I^-(y)$. In general, Alexandrov topology is coarser than the manifold topology, which is not difficult to see. The following shows that a chronal isomorphism preserves the underlying Alexandrov topologies of M and M'.

Proposition 4.33. Let $\phi : (M, g) \to (M', g')$ be a chronal isomorphism. Then the Alexandrov topologies on M and M' are homeomorphic.

Proof. Let $x, y \in M$. Then

$$\begin{aligned} \phi(I^+(x) \cap I^-(y)) &= \phi(I^+(x)) \cap \phi(I^-(y)) \\ &= I^+(\phi(x)) \cap I^-(\phi(y)), \end{aligned} \tag{4.26}$$

using Proposition 4.31. Thus, we can write $\phi(I^+(x) \cap I^-(y)) = I^+(z) \cap I^-(w)$, where $\phi(x) = z$ and $\phi(y) = w$; $w, z \in M'$. It follows that the image of a basic open set in the Alexandrov topology on M is a basic open set of Alexandrov topology on M'. Conversely, every Alexandrov basic open set in M' has a pre-image of the same type. $\square$

The following result shows that a chronal isomorphism carries the causal continuity (and similar causality conditions such as the strong causality) of M into the causal continuity of M'.

Theorem 4.34. Let $\phi : (M, g) \to (M', g')$ be a chronal isomorphism. Then (M, g) is causally continuous if and only if (M', g') is causally continuous.

Proof. Suppose M is reflecting. Suppose $I^+(z) \subseteq I^+(w)$ with z, w in M' and $z = \phi(x)$ and $w = \phi(y)$ for $x, y \in M$. Then, Proposition 4.32 gives $\phi(I^+(x)) \subset \phi(I^+(y))$, which gives $I^+(x) \subset I^+(y)$. But M being reflecting, we have $I^-(y) \subset I^-(x)$. This again gives, applying ϕ on both sides and using Proposition 4.32, that $I^-(w) \subset I^-(z)$ establishing the reflectingness of M'. The converse is seen similarly. Thus, we see that M is reflecting if and only if M' is.

Next, we show that a similar result holds for distinguishingness. Since M is causally continuous if and only if it is reflecting and distinguishing, the result follows. Let M be distinguishing and $z, w \in M'$ with $I^+(z) = I^+(w)$. Since ϕ is one-one and onto, we have $z = \phi(x)$ and $w = \phi(y)$ for some $x, y \in M$. Now $I^+(z) = I^+(w)$ gives $I^+(\phi(x)) = I^+(\phi(y))$, which gives by Proposition 4.32 $\phi(I^+(x)) = \phi(I^+(y))$. But ϕ is injective, so this gives $I^+(x) = I^+(y)$ or $x = y$ by the hypothesis. So $z = w$. In the same manner, we have $I^-(z) = I^-(w)$ implies $z = w$. $\square$

As shown by Proposition 4.33, the Alexandrov topologies of M and M' are equivalent in general whenever there is a chronal isomorphisms between the two. The same is not true, however, for the manifold topologies of M and M' and as shown by Malament (1977), distinguishingness is a necessary condition for this purpose.

Theorem 4.35. Let (M, g) and (M', g') be distinguishing space-times and $\phi : M \to M'$ be a chronal isomorphism. Then ϕ is a homeomorphism.

(By the results of Hawking, King and McCarthy (1976), ϕ is also a smooth conformal isometry).

Malament also showed by means of an example that the condition of future and past distinguishingness in the above theorem cannot be relaxed to either future distinguishingness or the past distinguishingness alone. However, as shown by Theorem 4.34, it is not necessary to assume the distinguishingness of both M and M' but this requirement can be relaxed to distinguishingness of only one of the space-times. Even though the chronal isomorphism preserves most of the causality conditions and the Alexandrov topology as shown above, one would say that the causal structure is really preserved when the metric tensors of M and M' are the same up to a conformal factor. From this point of view, Theorem 4.35 implies that a chronal isomorphism preserves causal structure when either M or M' is distinguishing. Then there is a smooth non-vanishing map $\Omega : M' \to \mathbb{R}$ such that $\phi_*(g) = \Omega^2 g'$, that is, ϕ is a smooth conformal isometry and the causal structures of M and M' are identical.

Hawking, King and McCarthy (1976) defined the path topology on a space-time as the finest topology which induces on all continuous timelike curves the same topology as given by the usual manifold topology. That is, $A \subseteq M$ is open in the path topology if and only if given any continuous timelike curve $\gamma : I \to M$, there is a standard manifold topology open set O such that $\gamma[I] \cap A = \gamma[I] \cap O$. They argue that whereas the usual manifold topology has no direct physical relevance, the path topology has several desirable features. In particular, it is motivated by the sense in which we experience continuity along all future directed continuous timelike curves. As pointed out by Malament (1977), such a topological structure of M can be recovered from its causal structure in the sense that when M is distinguishing and ϕ is a causal automorphism on M, then ϕ and ϕ^{-1} preserve future directed continuous timelike curves. Then, ϕ is a homeomorphism and a smooth conformal isometry according to Theorem 4.35. Another implication of ϕ being a conformal isometry is that the path topology also codes the differential and conformal structures of the space-time.

As noted earlier, the strong causality of a space-time is equivalent to the Hausdorff property of the Alexandrov topology of M. Thus, if ϕ is a chronal isomorphism from a strongly causal space-time (M, g) to any other space-time (M', g'), then Proposition 4.33 implies that the Alexandrov topology of (M', g') is Hausdorff, implying the strong causality of (M', g'). Now we can show that the global hyperbolicity of M is invariant under a chronal isomorphism.

Theorem 4.36. Let $\phi : (M, g) \to (M', g')$ be a chronal isomorphism. Then (M, g) is globally hyperbolic if and only if (M', g') is globally hyperbolic.

Proof. Suppose M is globally hyperbolic, then M' must be strongly causal as seen above. To show the compactness of $J^+(r) \cap J^-(s)$ for all $r, s \in M'$, first we note that there are p, q in M such that $\phi(p) = r, \phi(q) = s$ which follows from definition of ϕ. Now, strong causality implies distinguishingness, so ϕ is a homeomorphism by Theorem 4.35. Therefore, to deduce the compactness of $J^+(\phi(p)) \cap J^-(\phi(q))$, it is enough to show that this set equals $\phi[J^+(p) \cap J^-(q)]$ or $\phi[J^+(p)] \cap \phi[J^-(q)]$. Thus, we need to show $\phi(J^+(p)) = J^+(\phi(p))$. Suppose now $z = \phi(x) \in \phi((J^+(p))$, then there is a future directed non-spacelike curve from p to x. If this curve is timelike, it follows from Proposition 4.32 that $z = \phi(x) \in J^+(\phi(p))$. If this is a null geodesic, then also there is a null geodesic from $\phi(p)$ to $\phi(x)$ by a result of Hawking, King and McCarthy (1976). This implies that $\phi(J^+(p)) \subset J^+(\phi(p))$. The other part also follows similarly. □

As mentioned in the earlier part of this section, chronal automorphisms of the Minkowski space-time were studied by Alexandrov and Zeeman to show the interesting result that these maps must be orthochronous Lorentz transformations, up to translations, and dilatations. Lester (1984) also studied the chronal automorphisms of deSitter and Einstein cylinder space-times. The basic idea here is to show that chronal automorphisms preserve the null relationships between events in both directions. An important lemma used for this purpose by both Zeeman (1967) and Lester (1984) is available only for the exact solutions mentioned above. We give below a more general proof of this result. First, a note on the notation. We say that $x \to y$ if and only if $x < y$ but $x \not\ll y$, that is, $x < y$ and further $y \in J^+(x) - I^+(x)$.

Proposition 4.37. Let M be a causally simple space-time and x, y be distinct points of M. Then $x \to y$ if and only if $x \not\ll y$, and for all z, $z \ll x$ implies $z \ll y$.

Proof. Firstly, $x \to y$ implies $x < y$ and $x \not\ll y$. Now, let z be such that $z \ll x$. Then, $z \ll x < y$ implies, as noted in Section 4.1, that $z \ll y$. To see the converse, suppose $x \not\to y$ and $x \not\ll y$. Since M is causally simple, this gives, $y \notin J^+(x) \cup J^-(x) = \overline{I^+(x)} \cup \overline{I^-(x)}$ and so y is in the spacelike region of x. Hence $x \notin \overline{I^+(y)} \cup \overline{I^-(y)}$, but x is in the complement of this closed set. Let O be an open neighbourhood of x in this complement and let $z \in O \cap I^-(x)$. Then $z \notin \overline{I^+(y)} \cup \overline{I^-(y)}$ and hence $z \not\ll y$. But we have $z \ll x$. Thus, the negation of $x \to y$ implies the negation of the condition of the proposition, proving the result. □

4.6 Causal functions

Given a space-time (M, g), it is possible to introduce a notion of size or volume for subsets of M and in particular for the future and past of an event.

Using this notion we define here causal functions on M and show that functional properties such as one-oneness and monotonicity, and limiting behaviour of these functions along timelike curves, characterize important causal features of M such as global hyperbolicity and other causality conditions. A description of the boundary or ideal point structure of M is also obtained in terms of these functions. Thus, it turns out that causal functions have several useful applications as far as the overall causal structure of M is concerned and one obtains a unified description for the same. The significance of the alternative functional approach given here lies essentially in the unified and simple picture of causal structure which emerges when causal functions are employed.

To get an idea of the procedure used here, we begin by defining a finitely subadditive set function σ on M (Vyas and Joshi, 1983), the main requirement being that σ assigns a finite volume to M. Since the essential purpose is to assign numerical values to space-time sets, we first start with a rather general definition.

Let $\sigma : P(M) \to \mathbb{R}$ be a finitely subadditive set function, where $P(M)$ is the collection of all subsets of M, satisfying the following:

(i) $\sigma(\emptyset) = 0$, $\sigma(M) = 1$, for any open set $G \in P(M)$, $\sigma(G) > 0$;
(ii) $\sigma(A) = \sigma(\bar{A})$, for any $A \in P(M)$;
(iii) $A \subseteq B$ implies $\sigma(A) \leq \sigma(B)$.

Then the function $f^+ : M \to \mathbb{R}$, defined by

$$f^+(x) = \sigma(I^+(x)) \quad \text{for all} \quad x \in M,$$

is called a *future causal function* on M. The *past causal functions* are defined dually.

Vyas and Joshi (1983) pointed out, however, that the above criterion is too broad in that it allows for causal functions which have either unreasonable physical properties, or which contain very little information. To have an insight into what type of functions are possible as allowed by this definition, consider the set function ν given below which is defined only using the second countability of the manifold topology of M.

Let $\{U_n\}$ be a countable ordered basis of the topology of M. For any set $A \subseteq M$, we define

$$\begin{aligned} \nu_n(A) &= 1, \qquad \text{if} \qquad U_n \cap A \neq \emptyset, \\ \nu_n(A) &= 0, \qquad \text{if} \qquad U_n \cap A = \emptyset, \end{aligned}$$

and a set function ν by

$$\nu(A) = \sum_1^\infty 2^{-n}\nu_n(A).$$

It is easy to verify that ν assigns a finite measure to M, that is, $\nu(M) = 1$ and also all the properties of causal functions defined above are satisfied. Here the pair $f^{\pm}$ of causal functions is defined by

$$f^{\pm}(x) = \nu(I^{\pm}(x)),$$

for all $x \in M$. We can now prove the following.

Proposition 4.38. The functions $f^{\pm}$ are discontinuous on M.

Proof. Let $x, y \in M$ be any two points on a timelike curve. Then there exists a basic open set U_k such that $I^+(x) \cap U_k \neq \emptyset$ and $I^+(y) \cap U_k = \emptyset$. Hence $\nu(I^+(x))$ will contain the term 2^{-k} in its expression from the geometric series $\sum_1^\infty 2^{-n}$, whereas the same term will be absent from the expression for $\nu(I^+(y))$. It follows that f^+ jumps while we go from x to y even if x and y are arbitrarily nearer. Hence f^+ is discontinuous on M. □

In physical problems continuity is an essential property of a useful function and hence we should always like to choose causal functions with reasonable properties, avoiding the pathological features such as the above. Thus, one must be somewhat restrictive in the determination of the allowable class of causal functions, basically to avoid the possibilities such as $\sigma(A) = 1$ for all open sets $A \subseteq M$, which is allowed by the definition above.

There are two main alternatives here. The first one is to use the volume functions due to Geroch (1970a). Here one selects an arbitrary positive volume element dV on M such that the total volume of M is equal to unity:

$$\int_M dV = 1. \tag{4.27}$$

Then the *causal function* h^+ is defined as

$$h^+(x) = \int_{I^+(x)} dV > 0, \tag{4.28}$$

and h^- is defined dually. Such a volume element on M always exists; M being paracompact it is triangulable and one chooses a volume element on each of the four-simplices of M such that the volume of the first simplex is 1/2, that of the second is 1/4, and so on.

Another similar way to obtain $h^{\pm}$ is to use a finite measure on M (Hawking and Ellis, 1973), as below. Let $\{u_\alpha, \phi_\alpha\}$ be the countable atlas on M such that the closure of $\phi_\alpha(u_\alpha)$ is compact in $\mathbb{R}^4$. Let μ_0 be the natural Euclidian measure on $\mathbb{R}^4$ and let f_α be a partition of unity for the atlas given above. Then μ is defined in the following way

$$\mu(A) = \left(\sum_\alpha f_\alpha 2^{-\alpha}\{\mu_0[\phi_\alpha(u_\alpha)]\}^{-1}\phi^*_\alpha\mu_0\right)(A). \tag{4.29}$$

It is easy to check that $\mu(M) = 1$, and so on, and that μ has various desirable properties.

The causal functions are then defined by

$$h^\pm(x) = \mu(I^\pm(x)). \tag{4.30}$$

In fact, it is always possible to find an additive measure μ on M such that the volume of each open set O is positive and $\mu(M) = 1$. There are many such measures possible and they need not be directly related to the Lorentzian volume element $\sqrt{-g}\, d^4x$ on M. From now on, we shall always assume that such a choice has been made, which is of course not unique by definition.

We now consider the behaviour of causal functions on certain space-time models. The values of h are intuitively the 'size' of futures and pasts respectively (Fig. 32). In Minkowski space-time, if γ is any future directed timelike curve, h^+ is strictly decreasing in future. In this case, if γ is inextendible, h^+ will approach the value zero. Similarly, h^- is strictly increasing and could go to unity in future as the parameter along the curve tends to infinity. We note that for all future directed timelike curves, h^- will increase but need not converge to unity. For example, consider the timelike hyperbola

$$t = \sinh\lambda, \quad x = \cosh\lambda, \quad y, z = 0, \qquad \lambda \in \mathbb{R}.$$

If we move into the future along this curve, its past null cone approaches the null surface $x = t$. Hence h^- will not tend to unity.

Next, consider the two-dimensional Einstein cylinder (see Fig. 24) described by the space-time metric

$$ds^2 = (\cosh t - 1)^2(dt^2 - dx^2) + dtdx. \tag{4.31}$$

Since a point is cut out as shown, the space-time is causal. If γ is any future directed timelike curve, h^+ is decreasing along the same, but along the null geodesic given by $t = 0$ it has a constant value $1/2$. We note that the space-time is not future distinguishing here. By removing a strip we can make it past distinguishing and it can be seen that on $t = 0$ line h^- is not constant but is in fact monotone. A similar example is the

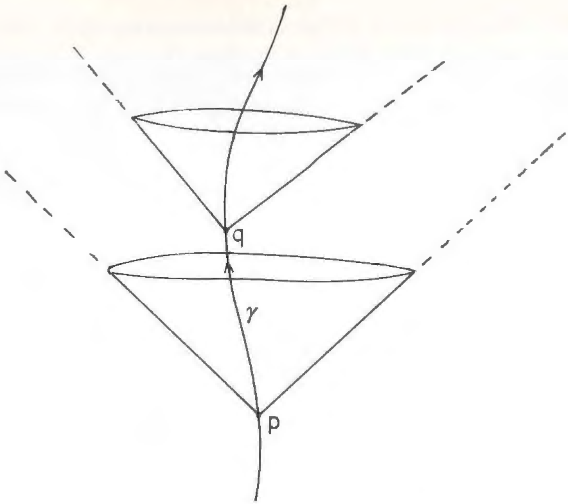

Fig. 32 Along a future directed timelike curve γ, the event q lies in the future of p and $I^+(q)$ is strictly contained in $I^+(p)$. Then, for a causal space-time, we must have $f^+(q) < f^+(p)$.

Misner space which is obtained when the metric on the cylinder is changed to $ds^2 = -(1/t)dt^2 + td\theta^2$ where $0 \leq \theta \leq 2\pi$ (Misner, 1967). There is a singularity at $t = 0$ in the metric and the coordinate t has range $0 < t < \infty$. However, one could extend the metric to include the $t < 0$ region, which is seen to admit closed timelike curves and $h^\pm$ are constant along these lines. There are no closed timelike lines when $t > 0$. One family of null geodesics is represented by vertical lines along which causal functions are monotonic. The other family spirals round and round as $t = 0$ is approached but never actually crosses this surface and these geodesics have finite affine length. Along such curves h^+ is decreasing without going to zero but tending to a number k with $0 < k < 1$. This behaviour brings out certain general features for causal functions as will be shown by results in this section.

Now we show that various important causal structure ingredients are characterized in terms of properties of the causal functions. The following is useful later.

Proposition 4.39 . If $x, y \in M$ are such that $I^+(x) \subset I^+(y)$ but $I^+(x) \neq I^+(y)$, then $\mu(I^+(x)) < \mu(I^+(y))$.

Proof. It is sufficient to show that the non-empty set $I^+(y) - I^+(x)$ contains an open set. For that, let $z \in I^+(y) - I^+(x)$. Therefore, $I^-(z) \cap I^+(x)$ must be empty, otherwise $z \gg x$. But $I^-(z) \cap I^+(y) \neq \emptyset$ and open because we have $z \gg y$. Thus $I^-(z) \cap I^+(y)$ is a non-empty open set contained in $I^+(y) - I^+(x)$. $\square$

Now we can characterize the chronology of M in terms of behaviour of causal functions.

Proposition 4.40 M is chronological if and only if $h^\pm$ are strictly monotone along each timelike curve.

Proof. Assume M to be chronological. Let γ be any future-directed timelike curve. We shall prove that h^+ is strictly decreasing along γ. Since γ is not closed, if we take $x, y \in \gamma$ with $x < y$ then it follows that $I^+(y)$ is proper subset of $I^+(x)$. Then $\mu(I^+(y)) < \mu(I^+(x))$ and hence by definition $h^+(y) < h^+(x)$. Conversely, the hypothesis of strict decreasing nature of h^+ along all timelike curves will keep M chronological. Because if γ is some closed timelike curve, then there are points $x, y \in \gamma$ with $x \ll y$ and $y \ll x$ which will give $h^+(y) < h^+(x)$ and $h^+(x) < h^+(y)$, which is a contradiction. $\square$

We discussed the behaviour of causal functions in a distinguishing space-time in the examples stated above. The distinguishingness of a space-time is characterized in terms of causal functions as below.

Proposition 4.41. For a space-time M, the following are equivalent: (i) M is future (past) distinguishing, (ii) h^+ (respectively h^-) is injective along all future (past)-directed non-spacelike curves, and (iii) h^+ (respectively h^-) is strictly decreasing (increasing) along all future (past)-directed non-spacelike curves.

Proof. To prove that (i) implies (ii), suppose M is future distinguishing, that is, $x \neq y$ implies $I^+(x) \neq I^+(y)$. Let γ be any future-directed non-spacelike curve. Let $x, y \in \gamma$ be such that $x < y$ but $x \neq y$. Now $x < y$ implies $I^+(x) \supseteq I^+(y)$. Now we already have $I^+(x) \neq I^+(y)$ so $I^+(y)$ is a proper subset of $I^+(x)$. This gives $h^+(x) > h^+(y)$ in the same manner as in Proposition 4.40. Thus, different points of γ have different images. Hence h^+ is injective along γ.

Next, to prove (ii) implies (iii), assume that h^+ is injective on any future-directed non-spacelike curve γ. Let $x, y \in \gamma$ with $x \neq y$. Without loss of generality we can take $x < y$, so we have $I^+(x) \supseteq I^+(y)$ giving $h^+(x) \geq h^+(y)$. But h^+ is injective on γ and $x \neq y$, so we get $h^+(x) > h^+(y)$. This proves the strict decreasing nature of h^+ along γ.

Finally, to prove that (iii) implies (i), let h^+ be strictly decreasing along all future-directed non-spacelike curves. Now, if possible assume that distinguishability is violated in M. So there are points $x, y \in M$ with $x \neq y$ but $I^+(x) = I^+(y)$. Now the last equality implies that x and y cannot lie in the spacelike region of each other (we say that q is in the spacelike region of p if $q \notin \overline{I^+(p)} \cup \overline{I^-(p)}$. Therefore, without loss of generality we have

$$x < y \quad \text{or} \quad y \in \dot{I}^+(x) - \{x\}.$$

Now $x < y$ would imply $h^+(y) < h^+(x)$ because of our hypothesis. But this is not possible since $I^+(x) = I^+(y)$ tells that $h^+(y) = h^+(x)$. Further if

$$y \in \dot{I}^+(x) - \{x\},$$

then taking $\{x\} = S$, a closed set, we have a null geodesic in $\overline{I^+(x)}$ whose future end point is y by Theorem 3.20 of Penrose (1972). Clearly this null geodesic is non-degenerate and so there is a point in the past of y on this geodesic. That is, there is a point $z \in \dot{I}^+(x)$ with $z < y$. So it follows that

$$h^+(z) > h^+(y). \tag{4.32}$$

Again $z \in \dot{I}^+(x)$ and so $I^+(z) \subseteq I^+(x)$, which says that

$$\mu(I^+(z)) \leq \mu(I^+(x)),$$

which implies $h^+(z) \leq h^+(x)$. Combining this with eqn (4.32) we get $h^+(y) < h^+(x)$. This is not possible because $I^+(x) = I^+(y)$ gives us $h^+(x) = h^+(y)$. Thus, we must have $I^+(x) = I^+(y)$ together with $x = y$ only. Hence M is future distinguishing. □

The non-occurrence of closed non-spacelike curves in the space-time, that is, causality is an important condition on M. The well-known singularity theorems generally assume the same. This condition is physically stronger as compared to chronology because here not only the material particles but photons also are not allowed to go back into their past. Geroch (1970a) noted that the characterization of causality in terms of 'volume functions' is somewhat complicated. We search here for conditions on causal functions which can make M causal. Since causality is stronger than chronology, $h^{\pm}$ must be strictly monotone along each timelike curve, by Proposition 4.40. At the same time, naturally, we must demand some more conditions on M. Now let $p, q \in M$ be joined by some non-spacelike curve γ. Then, either γ could be a timelike curve or it must be a null geodesic. It follows that our condition must be connected with null geodesics only. We do not demand monotonicity of $h^{\pm}$ along all null curves, because this, together with monotonicity along all timelike curves, will imply distinguishingness by Proposition 4.41. Thus, the condition on null trajectories must be lighter than monotonicity. Also, we may not expect constant nature along all null geodesics because in that case closed null geodesics might arise. Thus we are directed to keep the condition of non-constancy, that is, there are at least two points on each null geodesic where the future (past) causal function values differ. Any null geodesic is either complete or incomplete. Now, if γ is a closed null geodesic, incomplete in future, then a variation of γ can yield a closed timelike curve (Hawking and Ellis, 1973). Hence we shall impose the condition of non-constancy along all complete geodesics only. Then we can prove the following.

Proposition 4.42. A space-time M satisfies the condition of causality provided (i) $h^{\pm}$ are monotonic along all timelike curves, and (ii) $h^{\pm}$ are non-constant along all complete null geodesics.

Proof. If possible, suppose (i) and (ii) are satisfied and there is a closed non-spacelike curve γ in M. Then, as mentioned above, the three alternatives available are: (a) γ is a closed timelike curve, (b) γ is a closed complete null geodesics, (c) γ is a closed incomplete null geodesics. Now, (a) is immediately ruled out because then $h^{\pm}$ are constant along γ, contradicting (i). Similarly (b) is also ruled out by (ii). Now, for (c), we can apply the result mentioned above to convert γ into a closed timelike curve on which $h^{\pm}$ should be constant, contradicting the hypothesis. $\square$

We note that the converse in the above proposition can be shown to be holding under physically reasonable assumptions of an energy condition ($R_{ij}K^iK^j \geq 0$ for every null vector K^i) and the generic condition (that is, every null geodesic contains a point at which $K_{[i}R_{j]el[m}K_{n]}K^eK^l$ is non-zero, where K is the tangent vector.) If causality holds, clearly condition (i) holds by Proposition 4.40. Next, suppose h^+ is constant along some complete null geodesic γ. Let $x, y \in \gamma$ be any two different points, then $I^+(x) \subseteq I^+(y)$. But constancy of h^+ implies strict inclusion does not hold and hence $I^+(x) = I^+(y)$. But the energy condition and the generic condition imply that γ must contain a pair of conjugate points which are timelike connected (Hawking and Ellis, 1973). Thus we have $x, y \in \gamma$, $x \ll y$ and $I^+(x) = I^+(y)$, implying $x \in I^+(x)$, violating the chronology, which is not possible.

Hawking and Sachs (1974) introduced the condition of causal continuity on a space-time and pointed out its physical significance. Essentially, causal continuity prevents abrupt changes in the future or past of a point p as p moves along a non-spacelike curve. They showed that M is reflecting if and only if $h^{\pm}$ are continuous in M (Fig. 33).

Thus using Proposition 4.40, we can provide the following characterization for causal continuity.

Proposition 4.43. M is causally continuous if and only if $h^{\pm}$ are continuous and monotonic along all non-spacelike curves.

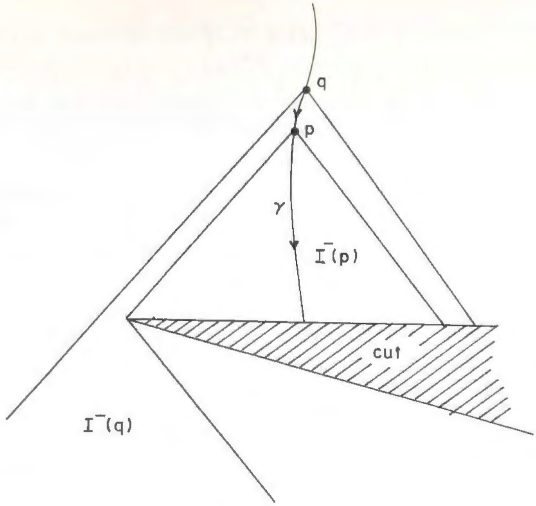

Fig. 33 The function h^- is not continuous along the non-spacelike curve γ and its value suddenly jumps in going from p to q. The space-time here is not reflecting.

A somewhat complicated proof of causal continuity implying stable causality was provided by Hawking and Sachs. We note that this follows in a simple manner from Proposition 4.43 as below. M being causally continuous, $-h^+$ is continuous and increasing along all future directed non-spacelike curves, providing a cosmic time function. Under this situation M must be stably causal (Sachs and Wu, 1977).

We now consider stable causality of M in terms of causal functions (Joshi, 1990), and recall from Section 4.2 that M is stably causal if and only if J_S^+ is a reflexive partial order on M, and that $J_S^+(x)$ is closed for all x in M. We define

$$h_s(x) = \mu(J_S^+(x)). \tag{4.33}$$

Theorem 4.44 The following are equivalent:

(a) M is stably causal

(b) $J_S^+(x) = J_S^+(y)$ implies $x = y$ for all $x, y \in M$.

(c) $h_s(x) = \mu(J_S^+(x))$ is decreasing along all future directed non-spacelike curves.

Proof. To see that (a) implies (b) suppose M is stably causal. Then, whenever $x \in J_S^+(y)$ and $y \in J_S^+(x)$, we have $x = y$. Suppose $J_S^+(x) = J_S^+(y)$. Then $x \in J_S^+(x)$ implies $x \in J_S^+(y)$ and, similarly, $y \in J_S^+(x)$. Hence by (a), $x = y$.

Conversely, suppose $x \in J_S^+(y), y \in J_S^+(x)$; we want to show that $x = y$ if (b) holds. Now, $x \in J_S^+(y)$ implies $J_S^+(x) \subseteq J_S^+(y)$ and $y \in J_S^+(x)$ implies $J_S^+(y) \subseteq J_S^+(x)$. Hence, $J_S^+(x) = J_S^+(y)$ and by (b), $x = y$. Assuming now (b) we derive (c). Let γ be a non-spacelike curve in (M, g) and $x, y \in \gamma$, such that $x < y, x \neq y$. Hence, $J_S^+(y) \subseteq J_S^+(x)$; however, $x \neq y$ implies $J_S^+(y) \subset J_S^+(x)$ without being equal. Now $J_S^+(x)$ being closed for all $x \in M$, taking interiors gives $Int J_S^+(y) \subset Int J_S^+(x)$ again without being equal. Consider now the set $A = Int J_S^+(x) - Int J_S^+(y)$. If $Int A = \emptyset$ then for any $q \in A$ and any neighbourhood N_q, $N_q \cap Int J_S^+(y) \neq \emptyset$ and hence $q \in \dot{J}_S^+(y)$. Then $q \in J_S^+(y)$. This implies $Int J_S^+(x) \subset J_S^+(y)$, or taking closure, $J_S^+(x) \subseteq J_S^+(y)$, which is not possible. Hence $Int A \neq \emptyset$ and $h_s(y) < h_s(x)$ (that is, h decreases along all non-spacelike curves).

Finally to show that (c) implies (b) let it be given that h_s is decreasing along all future directed non-spacelike curves. Suppose now for some x, y one has $J_S^+(x) = J_S^+(y)$ but $x \neq y$. This gives

$$h_s(x) = h_s(y), \qquad x \neq y. \tag{4.34}$$

Thus, x cannot be causally related to y, otherwise there will be contradiction. Now, $y \in J_S^+(y)$ implies $y \in J_S^+(x)$. Thus, $y \in \dot{J}_S^+(x)$ or $y \in Int J_S^+(x)$. In the former case, $I^+(Int J_S^+(x)) \subset Int J_S^+(x)$ implies there exists a non-degenerate null geodesic γ in $\dot{J}_S^+(x)$ with future end point y. Let $z \in \gamma$, then $J_S^+(z) \subseteq J_S^+(x)$; that is

$$h_s(z) \leq h_s(x). \tag{4.35}$$

Next, $z <_g y$ implies $h_s(y) < h_s(z)$. This, together with eqn (4.34) implies $h_s(y) < h_s(x)$, a contradiction with eqn (4.34). Similarly, if $y \in Int J_S^+(x)$, we can choose $z \in N_y \cap \gamma$ where N_y is a neighbourhood of y in $J_S^+(x)$ and γ is a past directed timelike curve with future end point y. A similar argument then gives the contradiction $h_s(y) < h_s(x)$. Hence we must have $x = y$ whenever $J_S^+(x) = J_S^+(y)$. $\square$

In order to characterize the strong causality condition, we use the sets $\uparrow I^-(x)$ defined by eqn (4.12). Let $\{z_n\} \subset I^-(x)$ and $z_n \to x$. Then

$$\begin{aligned}\mu(\uparrow I^-(x)) &= \mu\left(\cap_{n=1}^{\infty} I^+(z_n)\right) = \lim_n \mu(I^+(z_n)) \\ &= \lim_n h^+(z_n) \equiv h_m(x).\end{aligned} \tag{4.36}$$

It is possible to express the strong causality condition in terms of the sets $\uparrow I^-(x)$ as below (Penrose, 1972; Racz, 1987).

Lemma 4.45. M is strongly causal if the map $\uparrow I^-$ on M is a one-one map.

The functions $h_m(x)$ defined by eqn (4.36) can now be used as below.

Theorem 4.46. (M, g) is strongly causal if and only if the function h_m is decreasing along all future directed non-spacelike curves in M.

Proof. A sketch of the proof is given below. Let M be strongly causal and x, y be events on a non-spacelike curve γ with $x < y$, and $x \neq y$. Then, using Proposition 4.7 $\uparrow I^-(y) \subset \uparrow I^-(x)$ without being equal. The set $A = \uparrow I^-(x) - \uparrow I^-(y)$ has non-empty interior because otherwise it can be shown that $\uparrow I^-(x) = \uparrow I^-(y)$, which is a contradiction. Hence $h_m(y) < h_m(x)$. Conversely, if strong causality fails at $p \in M$, it is possible to use Theorem 4.31 of Penrose (1972) to show that the monotone nature of h_m is violated along an endless null geodesic γ through p. □

It may be noted that similar dual results hold, if we use the sets I^+ and $\downarrow I^+$, in terms of behaviour along past-directed trajectories.

The strongest form of a causality condition on M is global hyperbolicity, that is M admits a global Cauchy surface. The following characterizes global hyperbolicity in terms of causal functions, which requires the coexistence of disjoint properties like monotonicity, continuity, and suitable convergence of causal functions.

Proposition 4.47. A space-time M is globally hyperbolic if and only if $h^\pm$ are strictly monotonic and continuous along all timelike curves and $h^+ \to 0$ in the future along future inextendible timelike curves and $h^- \to 0$ in the past along past inextendible timelike curves.

Proof. If M is globally hyperbolic then monotonicity and continuity are satisfied because such a space-time is clearly distinguishing and causally continuous. Again to show $h^+ \to 0$, one can use the known argument that if some point q lay to the future of every point of a timelike curve γ, then γ would be future imprisoned in a compact set, which is not possible by global hyperbolicity (Geroch, 1970).

To prove the converse, define a function Φ by $\Phi(x) = h^+(x) - h^-(x)$ for all $x \in M$. Since h^+ and h^- are strictly decreasing, so is Φ along all future-directed timelike curves. Also Φ is continuous and its range is $(-1, 1)$. Now consider the set $S = \{x \in M \mid \Phi(x) = 0\}$, then S is achronal because of strict monotonicity of Φ and it is closed as can be seen using the continuity of Φ. Now let $p \in M$ be any point. If $\Phi(p) > 0$, then consider any future inextendible timelike curve γ from p. Since $h^+ \to 0$ on γ, it can be seen that Φ becomes negative somewhere on γ. Then continuity of Φ implies there must be $r \in \gamma$ with $\Phi(r) = 0$, that is, $p \in D^-(S)$. Similarly, if $\Phi(p) < 0$ then $p \in D^+(S)$. Thus, the domain of dependence of S is seen to be M; hence S is a Cauchy surface and M is globally hyperbolic. □

We add a note on space-times which not only violate all of the different causality conditions mentioned above but which are *totally vicious* in the sense that through every point of the space-time there passes a closed timelike curve. This amounts to the statement that there exists an event $p \in M$ such that $I^+(p) = M$ and $I^-(p) = M$. In this case, one can show that for all $q \in M$, $I^+(q) = M = I^-(q)$. Then it is easily seen that M is totally vicious if and only if the range of $h^\pm$ is $\{1\}$. When M is totally vicious, clearly we have $\mu(I^\pm(x)) = \mu(M) = 1$ for any $x \in M$. On the other hand, the range of $h^\pm$ being $\{1\}$ is possible only if for some $x \in M$, $I^+(x) = M = I^-(x)$.

Next, we provide here a description of the ideal points boundary of the space-time (that is, points at infinity and singularities) in terms of causal functions. Though the points at infinity and singularities in a space-time are not regular points for M, they can be attached to M as an additional boundary. Such a boundary construction including both the points at infinity and singularities was provided by Geroch, Kronherimer, and Penrose (1972), based only on the causal structure of space-time. We shall classify here this newly attached causal boundary or ideal points using the properties of causal functions. We begin with a few definitions.

A non-empty subset P of M is said to be a *past set* if there is some $A \subset M$ such that $I^-(A) = P$. If P cannot be expressed as the union of two proper past subsets then it is called an *indecomposable past set* (IP). If P is an IP and if there is some $x \in M$ with $I^-(x) = P$, then P is known as a *proper* IP or PIP. If an IP set is not a PIP, then it is termed as *terminal* IP or TIP. The definitions of future sets, indecomposable future sets (IF), PIF and TIF are dual. For many of the statements and propositions the duals are taken for granted. Geroch,Kronheimer and Penrose (1972) proved that a set $P \subset M$ is an IP if and only if there is a future-directed timelike curve γ such that $I^-(\gamma) = P$. Now, let M^+ be the union of $\hat{M}$ and $\check{M}$, which are unions of all IPs and IFs in M respectively. Then avoiding duplication in M^+, one can define M^* as the quotient space M^+/R_h , where R_h is the intersection of all equivalence relations $R \subset M^+ \times M^+$ for which M^+/R is Hausdorff. In that case, M^* can be viewed as a space-time with boundary, $M \subset M^*$ and the topology of M can be looked upon as the induced topology of M^*. Throughout the discussion here M has been assumed to be distinguishing.

A point $x \in M$ is said to be a *regular point* if it is represented by a PIP or PIF. All other points in M^* are represented by TIPs or TIFs and are called the *ideal* or *boundary points* of M. A curve γ in M is taken as a continuous map of a general interval into M. Geroch, Kronheimer and Penrose (1972) proved the following result, which is useful here.

Lemma 4.48. Let $P = I^-(\gamma)$ be an IP with γ a future-directed timelike curve. Then P is a PIP if and only if γ has a future end point and P is a TIP if and only if γ is future inextendible without a future end point.

We now characterize the PIPs (and PIFs) using causal functions.

Proposition 4.49. Let γ be a future-directed timelike curve. Then $I^-(\gamma)$ is a PIP if and only if h^- attains its maximum value along γ.

Proof. If $I^-(\gamma)$ is a PIP then γ has a future end point p by the lemma above. It follows that $I^-(\gamma) = I^-(p)$. Now M being distinguishing, h^- is strictly increasing along γ by Proposition 4.41 and hence has a maximum value at the future end point p.

Conversely, let h^- attain its maximum $h^-(p)$ for some $p \in \gamma$. Now if p is not the future end point of γ, then there exists some $q \gg p$ with $q \in \gamma$ and we have $h^-(q) > h^-(p)$ by the strict increasing nature of h^-, which contradicts the hypothesis. □

The above characterization can be stated in the following manner also. For a future-directed timelike curve γ, $I^-(\gamma)$ is a PIP if and only if h^+ attains its minimum value along γ. It now follows that for such a curve, $I^-(\gamma)$ is a TIP if and only if either $h^+ \to 0$ along γ or $h^+ \to k$ with $0 < k < 1$ along γ with $h^+(p) \neq k$ for any $p \in \gamma$. As mentioned earlier, along future endless curves in globally hyperbolic space-times, $h^+ \to 0$, whereas the other situation will be realized when points have been amputed from the space-time and the timelike curve converges to such a cut without having a future end point. Thus, the causal boundary points in M^* are defined by precisely those timelike curves along which $h^\pm$ do not realize their extremum values.

We note that the ideal points in $M^* - M$ include both the points at infinity and singularities characterized by TIPs and TIFs of the space-time. A TIP is *non-singular* if there exists a timelike curve generating it which has an infinite length. All other ideal points are singularities of the space-time. Now let x be a future (past) ideal point in M^* defined by a TIP, say $I^-(\gamma)$ (a TIF, say $I^+(\lambda)$). Then x will be called a *future (past) 0-ideal point* if h^+ (respectively h^-) converges to zero along the curve and x will be called a *future (past) k-ideal point* if h^+ (respectively h^-) converges to k with $0 < k < 1$ along γ (respectively, λ).

Thus, causal functions classify all the ideal points for M into the following eight categories:

(a) future (past) 0-ideal points which are singularities,
(b) future (past) k-ideal points which are singularities.
(c) future (past) 0-ideal points at infinity,
(d) future (past) k-ideal points at infinity.

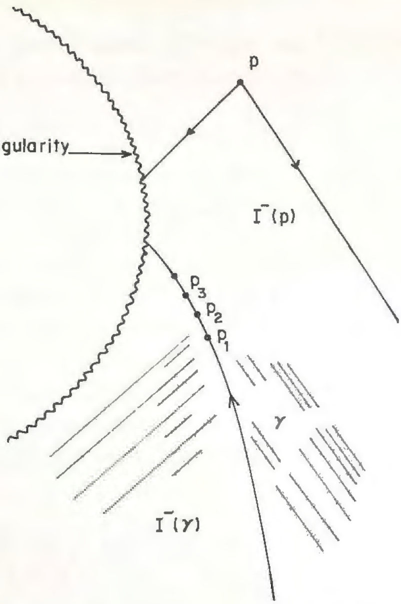

Fig. 34 The causal function h^+ does not vanish along γ as the curve falls into the singularity.

As mentioned earlier, endless timelike curves in a globally hyperbolic space-time define ideal points of type (c). Timelike curves falling into a Schwarzschild singularity provide an example for ideal points of type (a).

Finally, we show that the naked singularities as defined by Penrose (1974a) are all the boundary points of type (b). A well-known global problem in gravitation theory today is to decide whether the singularities are always hidden behind an event horizon. The question is whether naked or timelike singularities could also occur, which are defined as follows: let $x \in M^*$ be a TIP given by $I^-(\gamma)$. Then, x is called a *future naked singularity* if the future endless timelike curve γ has a future singular end point and there exists a point $p \in M$ such that $I^-(\gamma) \subset I^-(p)$. Now let x be a future singular point defined by a future endless timelike curve γ. As shown earlier, γ being a TIP, either $h^+ \to 0$ along γ or $h^+ \to k$ with $0 < k < 1$. Now consider any sequence $\{p_n\}$ on γ with $p_n \ll p_{n+1}$ for all n. Then $h^+(p_n)$ is strictly decreasing along this trajectory. Since $I^-(\gamma) \subset I^-(p)$, it follows that $p_n \ll p$ for all n. Thus $h^+(p_n)$ is bounded below by a non-zero fixed number $h^+(p)$, implying that $h^+ \not\to 0$ (Fig. 34). This shows that x is an ideal point of type (b).

4.7 Asymptotic flatness and light cone cuts of infinity

Consider a spherically symmetric star with a vacuum outside where the metric satisfies the empty space Einstein equations $R_{ij} = 0$ and is given by the Schwarzschild solution. In this case, far away from the star, as $r \to \infty$, the components g_{ij} tend to their Minkowskian values. Such isolated systems and the behaviour of gravitational field and metric components far away from the star at infinity are of considerable interest in general relativity. Even though the actual universe is not asymptotically flat as there would be matter present at all distances, such an approximation is quite useful in order to model the geometry of an individual star and to study its gravitational field.

It was pointed out by the studies of Bondi, van der Burg and Metzner (1962) and Sachs (1962) that the characteristic or null surfaces play a very important role in understanding the asymptotic properties of gravitational fields for such isolated systems. They used characteristic surfaces to study the metric components and curvature tensor properties in the asymptotic limit. Consider, for example, a burst of gravitational radiation coming out of the star with the initial wave front at a spacelike surface Σ_1 and let Σ_2 be a spacelike surface after the final burst has been emitted by the star (Fig. 35). However, these spacelike surfaces are not adequate to measure the total energy emitted in the form of waves because all the emission in the form of gravitational waves will intersect Σ_2 as well and so the total energy density as measured on Σ_1 and Σ_2 will be the same. On the other hand, if one considers characteristic surfaces, say the future light cone at the star at Σ_1, and at the later instance Σ_2, then the data on such null hypersurfaces should record the difference in the energy emitted in the form of the gravitational waves. As we shall show in the present section, the consideration of such null hypersurfaces throw much light on the structure of asymptotically flat space-times.

As pointed out above, in order to study the asymptotic behaviour in Schwarzschild geometry, one would take the limit as $r \to \infty$. However, to study the asymptotic behaviour in general, using a coordinate system is not very helpful. Because of the equivalence of all allowable coordinate systems in general relativity, any conclusion cannot be accepted until it is shown to be valid in a coordinate independent manner. Thus, one would like to introduce the notion of infinity for a space-time in a coordinate independent manner and then the behaviour of fields can be studied in the vicinity of the same.

Such a construction of infinity and the notion of asymptotic flatness for a general space-time were introduced by Penrose (1965, 1968) by means of the conformal compactification of a space-time. Here, the main idea is to attach a boundary to the space-time in such a manner that the properties

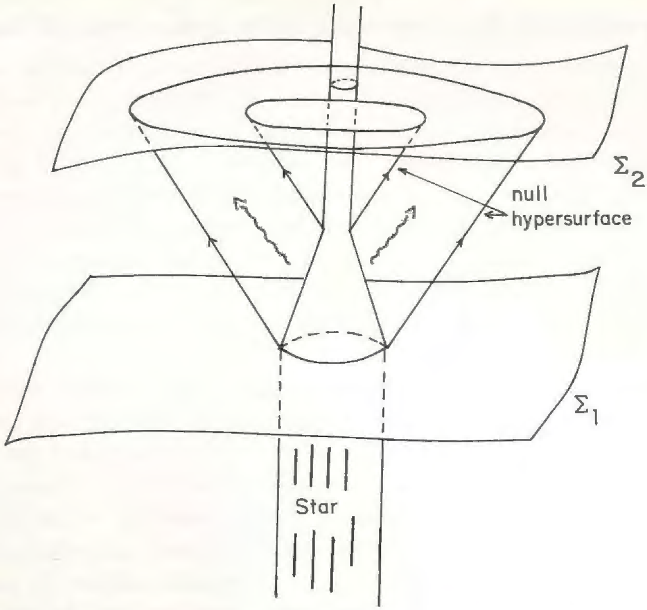

Fig. 35 Total mass-energy densities measured on the spacelike hypersurfaces Σ_1 and Σ_2 are the same. However, the characteristic data defined on the null surfaces shown will register the amount of gravitational radiation emitted by the star.

of the attached boundary must coincide with the geometric properties of the boundary $\mathcal{I}^+$ or $\mathcal{I}^-$ for the Minkowski space-time (Fig. 12) discussed in Section 3.1. A general space-time is then called asymptotically flat if it admits such a boundary attachment. Just as in the Minkowski space-time, one introduces here a conformal transformation Ω on the original space-time M so that $\Omega \to 0$ near infinity and the new unphysical space-time $(\overline{M}, \Omega^2 g_{ij})$ is compactified. In $(\overline{M}, \Omega^2 g_{ij})$ the boundary surface of $\overline{M}$ corresponds to the infinity of the space-time M.

Specifically, a space-time (M, g_{ij}) is called *asymptotically flat* if there exists a new, unphysical space-time $(\overline{M}, \bar{g})$ with a boundary $\mathcal{I}$ such that $\overline{M} - \mathcal{I}$ is diffeomorphic to M with $\Omega > 0$ on M and $\bar{g}_{ij} = \Omega^2 g_{ij}$ and the following are satisfied:

1. the new unphysical manifold $\overline{M}$ is smooth everywhere including the boundary;
2. the conformal factor Ω is smooth everywhere and $\Omega = 0$ on $\mathcal{I}$;
3. all the maximally extended null geodesics in $\overline{M}$ have a future and a past end point on $\mathcal{I}$;

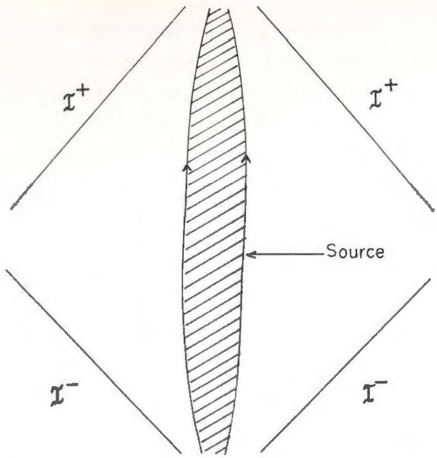

Fig. 36 An isolated source in an otherwise empty and asymptotically flat space-time universe.

4. there is a neighbourhood of $\mathcal{I}$ in M where g_{ij} satisfies the vacuum Einstein equations $R_{ij} = 0$.

Such a construction of conformal compactification for a space-time turns out to be particularly useful in order to study isolated sources in otherwise empty space-times (Fig. 36).

It is not difficult to see that the above properties are effectively the same as the properties of the future and past null infinities $\mathcal{I}^{\pm}$ constructed for the Minkowski space-time in Section 3.1. The surface $\mathcal{I}$ can be thought of as infinity in the sense that the affine parameter along every null geodesic in M grows unboundedly large near $\mathcal{I}$. The null geodesics of these two conformally related space-times are completely identical as point sets as we have pointed out earlier; however, the affine parameters along the geodesics in M and $\overline{M}$ are related as $d\bar{v} = \Omega^2 dv$. Thus, the affine parameter along the null geodesics in M must blow up near $\mathcal{I}^+$ in future, and similar behaviour holds near the past infinity. In fact, it is possible to show (Penrose 1968) that just as in the case of the Minkowski space-time, the vector

$$\frac{\partial \Omega}{\partial \bar{x}^i} \neq 0 \qquad \text{and} \qquad \bar{g}^{ij} \frac{\partial \Omega}{\partial \bar{x}^i} \frac{\partial \Omega}{\partial \bar{x}^j} = 0, \tag{4.37}$$

on $\mathcal{I}^+$ and also that the infinity $\mathcal{I}$ consists of two disjoint parts $\mathcal{I}^+$ and $\mathcal{I}^-$, each of which could have the topology of $\mathbb{R} \times S^2$. We note that the topology of M as a whole is that of $\mathbb{R}^4$ in case of an asymptotically flat space-time. It follows that $\mathcal{I}$ is a null hypersurface and so are $\mathcal{I}^+$ and $\mathcal{I}^-$.

It was pointed out by Penrose (1965) that the Weyl tensor of the metric g_{ij} vanishes at $\mathcal{I}$ and from this the sequential degeneracy or the peeling property of the Weyl tensor follows, namely that various components of the Weyl tensor vanish along the null geodesics as different powers of the affine parameter. Also, the conformal invariance of the equations for massless fields imply that the scattering problem for the massless particles in an asymptotically flat space-time reduces to the Cauchy problem in general relativity and what one needs to do is to study the properties of such fields locally near a null hypersurface which is the null infinity.

The definition of the asymptotically flat space-times given above does not assume any causality assumption on the space-time. It is, however, convenient to assume sometimes that the space-time is strongly causal, as it simplifies the considerations. In fact, it was pointed out by Newman and Clarke (1987) that if the topology of the future or past null infinities is to be $S^2 \times \mathbb{R}$, then in general one needs the additional assumption that not only M is strongly causal but the unphysical space-time $\overline{M}$ also must be strongly causal. This means that the unphysical metric has to be strongly causal even at $\mathcal{I}$. Further, rather than treating just the null infinity, it is possible to give a definition which gives a unified treatment for both the spatial and null infinities following the approach given by Ashtekar and Hansen (1978). For details of these generalizations we refer to Wald (1984). The definition above is quite stringent in that it assumes that every null geodesic has two end-points, in future and in past at $\mathcal{I}$. Though this is satisfied in the Minkowski space-time, such a property does not hold in space-times such as the Schwarzschild and Reissner–Nordström cases, which contain event horizons and a black hole region. The future directed null geodesics which enter the black hole must end up in the singularity at $r = 0$ and cannot have end point at $\mathcal{I}^+$. One would like to include these space-times in the general asymptotically flat class and hence we define a space-time M to be *weakly asymptotically simple and empty* (or a WASE space-time), if there exists an asymptotically flat space-time $\overline{M}$ in the above sense such that there is a neighbourhood of $\mathcal{I}$ in $\overline{M}$ which is isometric to an open set in M. This definition covers the Schwarzschild and the Reissner–Nordström cases and also the Kerr solutions (Hawking and Ellis, 1973).

In Sections 3.1 and 3.2 we introduced complex stereographic coordinates ζ and $\bar{\zeta}$ on the sphere for M and on $\mathcal{I}^+$, which are more suitable for the study of asymptotic structure of M in the neighbourhood of $\mathcal{I}^+$. As we noted above, when the topology of $\mathcal{I}^+$ is $S^2 \times \mathbb{R}$ for a general asymptotically space-time, these coordinates can again be used in a general situation. This provides a more convenient way to study $\mathcal{I}^+$ and the light cone cuts of null infinity as we will see here. The surface $\mathcal{I}^+$ is a three-dimensional null hypersurface which is generated by a two-parameter family of null geodesic

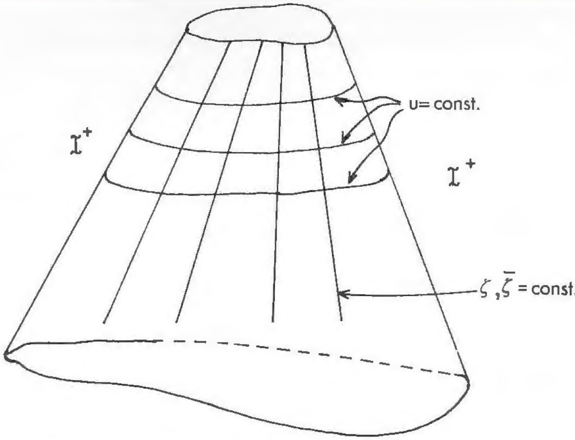

Fig. 37 Bondi coordinates at the future null infinity.

generators (similar statements hold for the past null infinity). Considering the topological structure of $\mathcal{I}^+$, one could choose two-dimensional non-intersecting spacelike surfaces, each diffeomorphic to S^2, so that a one-parameter family of such surfaces, parametrized by a variable u, provides a foliation for $\mathcal{I}^+$. Choosing $u = \text{const.}$ gives one of these surfaces on which we introduce the complex stereographic coordinates ζ and $\bar{\zeta}$ defined earlier. Thus, we have coordinates $(u, \zeta, \bar{\zeta})$ on $\mathcal{I}^+$, which are called the *Bondi coordinates* on $\mathcal{I}^+$ (Fig. 37). These coordinates are determined up to a conformal transformation on the sphere itself and up to a new choice of the family of sections $u = \text{const.}$

Having introduced the general framework for dealing with asymptotically flat space-times, the purpose of the rest of this section is to study the light cone cuts of $\mathcal{I}^+$ from an interior apex point in such a space-time. We point out that the knowledge of such cuts yields considerable information about the interior space-time. In particular, by an explicit consideration of light cone cuts for the Schwarzschild space-time, we show that it is possible to reconstruct the interior space-time metric up to a conformal factor from the knowledge of such cuts. A consideration for light cone cuts in Kerr–Newman space-times is also given, which incorporates the angular momentum as well as charge. For a general description of the theory of light cone cuts in an asymptotically flat space-time, we refer to the review by Kozameh and Newman (1986).

Let M be an asymptotically flat space-time and x^i be an arbitrary interior point. If N_x denotes the light cone at x then the intersection of this cone with $\mathcal{I}^+$, given by $N_x \cap \mathcal{I}^+$, is locally a two-surface which is uniquely labelled by the apex of the cone x^i. Using the Bondi coordinates on $\mathcal{I}^+$, the equation for these cuts can be given in the form

$$u = Z(x^i, \zeta, \bar{\zeta}).$$

Given the space-time metric up to a conformal factor it is possible, in principle, to integrate all the null geodesics in the space-time and also to obtain the full light cone at any point x. This light cone is locally surface forming, except for the conjugate points, and can be described by the equation

$$L(x^i, x^{i'}) = 0, \tag{4.38}$$

where x^i is the apex of the cone which is fixed, and $x^{i'}$ are points on the null surface. Fixing $x^{i'}$ the set of points x^is given by the above will form the light cone coming out from $x^{i'}$. In terms of the above function L, the light cone cuts $N_x \cap \mathcal{I}^+$ of $\mathcal{I}^+$ will be given by

$$L(x^i, \mathbf{x}^{\mathbf{i}'}) = \mathbf{0}, \tag{4.39}$$

where we have denoted the points confined to $\mathcal{I}^+$ by $\mathbf{x}^{\mathbf{i}'}$. Again, such cuts will be two-surfaces locally and using the Bondi coordinates on $\mathcal{I}^+$, we could solve for u in eqn (4.38) to write the cut function locally as

$$u = Z(x^i, \zeta, \bar{\zeta}). \tag{4.40}$$

For a fixed apex x in the space-time, eqn (4.39) describes the light cone cut C_x, so varying the interior point x generates a four-parameter family of cuts at $\mathcal{I}^+$. On the other hand, if we fix a point $(u, \zeta, \bar{\zeta})$ at $\mathcal{I}^+$, then the points x^i that satisfy eqn (4.39) form the past null cone of a point of $\mathcal{I}^+$. According to this later interpretation, $Z_{,i}(x, \zeta, \bar{\zeta})$ is a null vector and so it obeys the equation

$$g^{ij}(x) Z_{,i}(x, \zeta, \bar{\zeta}) Z_{,j}(x, \zeta, \bar{\zeta}) = 0. \tag{4.41}$$

Now, we assume that the cut function Z is given and we would like to reconstruct the interior space-time metric up to a conformal factor. After describing the basic technique involved, we apply this method in an explicit manner to the case of the Schwarzschild and the Kerr–Newman space-times (Joshi, Kozameh and Newman, 1983; Joshi and Newman, 1984). The basic reconstruction technique consists in applying the differential operators $\eth$ and $\bar{\eth}$ (Newman and Penrose, 1966; these correspond respectively to the differentiation with respect to ζ and $\bar{\zeta}$ in a suitably weighted manner), to eqn (4.40) several times until enough equations are obtained, so that the metric components could be written explicitly in terms of the gradient basis

$Z_{,i}$, $\eth Z_{,i}$,$\bar{\eth} Z_{,i}$, $\eth\bar{\eth} Z_{,i}$. For example, let us define a tetrad basis θ_i^a (a is the tetrad index with values $0, +, -, 1$) as below:

$$\theta_i^0 = Z_{,i}, \quad \theta_i^+ = \eth Z_{,i}, \quad \theta_i^- = \bar{\eth} Z_{,i}, \quad \theta_i^1 = \eth\bar{\eth} Z_{,i}. \tag{4.42}$$

In terms of such a basis, the tetrad components of the metric can be written as

$$g^{ab} = g^{ij}\theta_i^a\theta_j^b, \tag{4.43}$$

and eqn (4.40) reads in the new notation as

$$g^{ij}\theta_i^0\theta_j^0 = g^{00} = 0. \tag{4.44}$$

Applying the operators $\eth$ and $\bar{\eth}$ to g^{00} we obtain $g^{0+} = 0$ and $g^{0-} = 0$. Operating again with $\eth$ on this last relation, we obtain $g^{+-}/g^{01} = -1$. In this way the trivial tetrad components of the metric are obtained. For the rest of the components, further $\eth$ and $\bar{\eth}$ operations are required.

In order to get the remaining six components, the quantities $\eth^2 Z$ and $\bar{\eth}^2 Z$ are important and can be used for this purpose. We can write the expansion of $\eth^2 Z_{,i}$ in terms of the basis θ_i^a as

$$\eth^2 Z_{,i} = \Lambda_0\theta_i^0 + \Lambda_+\theta_i^+ + \Lambda_-\theta_i^- + \Lambda_1\theta_i^1 \equiv \Lambda_a\theta_i^a. \tag{4.45}$$

Since Z is known, θ_i^a and $\eth^2 Z_{,i}$ are also known quantities and so the above leads to an algebraic determination of the quantities Λ_as. It can be shown then (Joshi, Kozameh and Newman, 1983) that there is a one-one algebraic relationship between the quantities

$$q_{ab} \equiv (1/g^{01})(g^{++}, g^{--}, g^{+1}, g^{-1}, g^{11}, \eth g^{01}, \bar{\eth} g^{01}),$$

and the quantities λ_μ given by

$$\lambda_\mu \equiv (\Lambda_1, \bar{\Lambda}_1, \Lambda_+, \bar{\Lambda}_+\Lambda_-, \bar{\Lambda}_-, \Lambda_0\bar{\Lambda}_1 + \Lambda_0\Lambda_1).$$

This correspondence could be used in two ways, namely, to obtain the tetrad metric components q^{ab} if the quantities Λ_μ are known, which will be the case of interest here, and alternatively one could obtain Λ_μ if the data given is the tetrad components of the metric. Once the tetrad components of the metric are known, the coordinate components g^{ij} can be obtained using the relationship $g^{ij} = g^{ab}\theta_a^i\theta_b^j$, where the tetrad vectors satisfy $\theta_a^i\theta_j^a = \delta_j^i$.

We now study such light cone cuts in the Schwarzschild space-time. This is the simplest non-trivial solution of Einstein equations which is asymptotically flat and has considerable physical applications. Obtaining light cone cuts explicitly in the case of Schwarzschild geometry yields much insight into the general theory of light cone cuts of infinity outlined above, and provides an explicit demonstration for the same. Further, such a study generates explicit expressions for quantities such as the cut function Z, $\eth^2 Z$ and such others which have intrinsic importance from the point of view of the asymptotic structure of space-time as we shall discuss below.

Using the known Schwarzschild metric, the null geodesics in the space-time can be worked out and this provides a complete description of the light cone from an arbitrary apex in the space-time. The intersection of this cone with $\mathcal{I}^+$ generates the light cone cut. The non-trivial feature present here, as compared to the flat space-time, is the angular deflection of the of null rays due to curvature. Because of this factor, the cut function Z in this geometry is described by elliptic integrals and must be given parametrically as opposed to the flat space case where one gets elementary functions. In fact, a portion of the light cone cut (that generated by the null rays in the equatorial plane) for the flat space case was already given in Section 3.1 by the eqn (3.25), which is S^1 worth of null rays. To generate the full cut now, we can rotate the equatorial plane, using the spherical symmetry of the situation.

For this purpose, it is convenient to define a null vector l^a by

$$l^a = \frac{1}{\sqrt{2}}\left(1, \frac{\zeta+\bar{\zeta}}{1+\zeta\bar{\zeta}}, \frac{-i(\zeta-\bar{\zeta})}{1+\zeta\bar{\zeta}}, \frac{\zeta\bar{\zeta}-1}{1+\zeta\bar{\zeta}}\right). \tag{4.46}$$

As $\zeta, \bar{\zeta}$ move over S^2, this vector swips the null cone of directions. In Section 3.1 the situation considered was that of all null geodesics in the equatorial plane with the apex on the $\phi = 0$ axis. Now, let $\vec{n}_A = (1, 0, 0)$ and $\vec{n}_F = (\cos\phi, \sin\phi, 0)$ denote two unit vectors at the origin, pointing respectively to the origin and the final direction of the geodesic. We have $n_A . n_F = \cos\phi$. Performing now an arbitrary rigid rotation so that the apex moves to a new direction,

$$n'_A = (\sin\theta_0 \cos\phi_0, \sin\theta_0 \sin\phi_0, \cos\theta_0), \tag{4.47}$$

and the final point to

$$n'_F = (\sin\theta \cos\phi, \sin\theta \sin\phi, \cos\theta). \tag{4.48}$$

Then we have, $\vec{n}_A . \vec{n}_F = \vec{n}'_A . \vec{n}'_F$ which means

$$\cos\hat{\phi} = \cos\theta_0 \cos\theta + \sin\theta_0 \sin\theta \cos(\phi - \phi_0), \tag{4.49}$$

where $\hat{\phi}$ denotes the angle between the directions $\vec{n}_A$ and $\vec{n}_F$. Now, using the coordinates ζ and $\bar{\zeta}$ instead of θ and ϕ, the eqn (4.48) and the relation

$$(\sin\theta\cos\phi, \sin\theta\sin\phi, \cos\theta) = \left(\frac{\zeta+\bar{\zeta}}{1+\zeta\bar{\zeta}}, \frac{-i(\zeta-\bar{\zeta})}{1+\zeta\bar{\zeta}}, \frac{\zeta\bar{\zeta}-1}{1+\zeta\bar{\zeta}}\right)$$

we get

$$\cos\hat{\phi} = 1 - 2l^a\tilde{l}_a, \tag{4.50}$$

where $\tilde{l}_a$ corresponds to the values $(\zeta_0, \bar{\zeta}_0)$ in eqn (4.45) instead of the values $(\zeta, \bar{\zeta})$. Substituting this general value for the angle in the equation for the cut (3.25) in the equatorial plane gives

$$u = u_0 + \frac{1}{l_0}(l^a\tilde{l}_a) = Z(u_0, l_0, \zeta_0, \bar{\zeta}_0, \zeta, \bar{\zeta}). \tag{4.51}$$

This is the general equation of light cone cut of $\mathcal{I}^+$ in the Minkowski spacetime where the apex has the coordinates $(u_0, l_0, \zeta_0, \bar{\zeta}_0)$.

We now return to the Schwarzschild geometry. Without loss of generality we can restrict to the null geodesic family in the first sheet. Then, from eqns (3.44) in Section 3.2 we have

$$du = -\frac{b^2 dl}{2\sqrt{A}(1+\sqrt{A})}.$$

Integrating the above gives

$$u = u_0 - \frac{1}{2}\int_{l_0}^{l}\frac{b^2 dl'}{\sqrt{A}} + \frac{1}{2}\int_{l_0}^{l}\frac{b^2 dl'}{1+\sqrt{A}} = u(u_0, l_0, b, l). \tag{4.52}$$

It should be noted that the quantity A is a cubic and hence both the integrals appearing above are elliptic integrals. These could be rewritten in terms of the standard elliptic integrals of first, second, or third type. This is a consequence of the non-zero value for the mass parameter m in A. If $m = 0$ then the integration is elementary and the result coincides with the flat space situation.

Choosing the apex on the $\phi = 0$ axis, the remaining geodesic equation for the angle can be integrated to obtain

$$\phi = -\int_{l_0}^{l}\frac{b\,dl'}{\sqrt{A}} \equiv \Phi(l_0, l, b). \tag{4.53}$$

One could now pass to the limit $l = 0$ in the eqns (4.51) and (4.52) which corresponds to $r = \infty$ to obtain equations of the form

$$u = U(u_0, l_0, b), \qquad \hat{\phi} = \Phi(l_0, b). \tag{4.54}$$

By eliminating b from the above equations, the above implicitly defines a function

$$u = u(u_0, l_0, \hat{\phi}), \tag{4.55}$$

which yields the equatorial portion of the light cone cut of $\mathcal{I}^+$. In order to generate the full cut one can now adopt the same method as used in the Minkowski case discussed above using the spherical symmetry of the space-time. Thus we get

$$\cos\hat{\phi} = 1 - 2l_a(\zeta, \bar{\zeta})\tilde{l}^a(\zeta_0, \bar{\zeta}_0). \tag{4.56}$$

So, the full light cone cut at $\mathcal{I}^+$ can be written in the form

$$u = Z(u_0, l_0, \zeta_0, \bar{\zeta}_0, \zeta, \bar{\zeta}), \tag{4.57}$$

which involves the apex point in the space-time, and the coordinates $(\zeta, \bar{\zeta})$ at $\mathcal{I}^+$. As eqns (4.55) and (4.56) cannot be solved explicitly, we must give this cut function Z parametrically as

$$u = u_0 + \frac{1}{2}\int_0^{l_0} \frac{b^2 dl}{\sqrt{A}} - \frac{1}{2}\int_0^{l_0} \frac{b^2 dl}{1 + \sqrt{A}}, \tag{4.58}$$

$$\hat{\phi} \equiv \cos^{-1}(1 - 2l_a\tilde{l}^a) = \int_0^{l_0} \frac{bdl}{\sqrt{A}}. \tag{4.59}$$

As far as the second sheet is concerned, the calculation of the cut function and its parametric equations can be obtained in a similar manner. In this case, the integration is slightly more complicated as compared to the $\dot{l} < 0$ case in that first one has to integrate with $+\sqrt{A}$ for a fixed value of b up to the bounce point l_b given by the solution of $A(l_b) = 0$, after which we return to the first sheet $-\sqrt{A}$ and integrate up to $l = 0$. The resulting expressions are similar as above.

Now that the light cone cuts of $\mathcal{I}^+$ have been worked out explicitly for the Schwarzschild geometry, the main task in this connection is accomplished because by the general theory indicated above, the space-time metric is determined. However, for several reasons the calculation other way is important, namely given the cut function Z the construction of the metric up to a conformal factor. Firstly, it serves as a check on the general formalism described above and on the correctness of the cut function derived, which is an important quantity as far as the asymptotic structure of the space-time is concerned (see for example, the review by Newman and Tod, 1980). Secondly, in this process several other quantities are worked out which have intrinsic interest of their own. We refer to Joshi, Kozameh and Newman (1983) for details of the calculation. Given the cut function

Z for the Schwarzschild geometry, the tetrad basis vectors θ^i_a have been worked out there for this case, and also the quantities $\eth^2 Z$, $\eth^2 Z_{,i}$ and their complex conjugates are obtained. Given this data, the quantities λ_μ are worked out by expanding $\eth^2 Z_{,i}$ in terms of the tetrad vectors. Then, as discussed above, the tetrad components as well as the coordinate components of the metric tensor are worked out, completing the reconstruction of the interior metric g_{ij} up to a conformal factor.

We now consider the light cone cuts of null infinity for the Kerr-Newman space-time in some detail (Joshi and Newman, 1984). One again chooses an apex x^i_0 in the interior space-time and null geodesics are to be obtained starting at this apex and ending at $\mathcal{I}^+$. For this purpose, the affine parameter is eliminated from the eqns (3.58) of Section 3.4 and then integration gives the light cone from this apex in terms of the elliptic integrals (Carter, 1968)

$$\begin{aligned} \int_{\theta_0}^{\theta} \frac{d\theta}{\sqrt{\Theta}} &= \int_{r_0}^{r} \frac{dr}{\sqrt{R}} \\ \phi &= \int_{\theta_0}^{\theta} \frac{(L\sin^{-2}\theta - a)}{\sqrt{\Theta}} d\theta + \int_{r_0}^{r} \frac{a}{\Delta}\left(1 + \frac{P}{\sqrt{R}}\right) dr \\ u &= \int_{\theta_0}^{\theta} \frac{(aL - a^2\sin^2\theta)}{\sqrt{\Theta}} d\theta + \int_{r_0}^{r} \frac{r^2+a^2}{\Delta}\left(1 + \frac{P}{\sqrt{R}}\right) dr. \end{aligned} \tag{4.60}$$

Now, taking the limit in the above as $r \to \infty$ we obtain the light cone cuts of $\mathcal{I}^+$ in the parametric form when θ is replaced by θ_∞ and ϕ by ϕ_∞. Then the first of eqns (4.59) is viewed as the implicit version of

$$\theta_\infty = \theta_\infty(\theta_0, r_0, , L, K), \tag{4.61}$$

and the second equation in eqns (4.59), using eqn (4.60) has the form,

$$\phi_\infty = \phi_\infty(\theta_0, \phi_0, r_0, L, K). \tag{4.62}$$

Finally, the equation for u can be written, using the above equations, as

$$u = u(u_0, r_0, \theta_0, L, K). \tag{4.63}$$

However, eqns (4.60) and (4.61) show that L and K can be viewed as functions of $\theta_0, \phi_0, r_0, \theta_\infty, \phi_\infty$. Thus, by eliminating L and K from eqn (4.62) we obtain the light cone cut function in the usual form

$$u = Z(r_0, \theta_0, \phi_0, \theta_\infty, \phi_\infty).$$

In this case, since we are considering only the null geodesics from the given apex, there is a specific allowed range for the values of the parameters L and K as specified by Joshi and Newman (1984). We note that in the case of the parameters $a = e = 0$, the above description for the cut function reduces to the Schwarzschild situation.

It is clear from the discussion so far that the cut function Z contains a great deal of information on the interior space-time and in particular determines the conformal metric. It is possible that information that may be difficult to obtain directly from the conformal metric could be easily obtained from Z. For example, a direct differentiation of Z gives all the null geodesics of M and thus all null surfaces and the parametrized description of all the null cones. In fact, for the Kerr or Schwarzschild geometry, the asymptotic shear σ_{x^i} of the light cone of the point x^i is given by

$$\sigma_{x^i}(\zeta,\bar{\zeta}) = \eth^2 Z(x^i,\zeta,\bar{\zeta}). \tag{4.64}$$

This equation can be used to reformulate the Einstein equations in the following manner. Given the cut function Z for the above geometries, one calculates $\eth Z$, $\bar{\eth} Z$, and $\eth\bar{\eth} Z$, which can be considered as the parametric versions of

$$x^i = x^i(Z, \eth Z, \bar{\eth} Z, \eth\bar{\eth} Z, \zeta, \bar{\zeta}). \tag{4.65}$$

If one could eliminate now x^i between eqn (4.63) and its complex conjugate, and eqn (4.64), we obtain

$$\eth^2 Z = \Lambda(Z, \eth Z, \bar{\eth} Z, \eth\bar{\eth} Z, \zeta, \bar{\zeta}), \tag{4.66}$$

$$\bar{\eth}^2 Z = \bar{\Lambda}(Z, \eth Z, \bar{\eth} Z, \eth\bar{\eth} Z, \zeta, \bar{\zeta}). \tag{4.67}$$

The above serves as a pair of differential equations which determines $Z = Z(x^i,\zeta,\bar{\zeta})$. The general solution appears to depend on four arbitrary parameters which could be taken as the local interior coordinates x^is. From this point of view, the above could be viewed as the Einstein equations for the Kerr geometries. For further details on this approach, we refer to the review by Newman and Tod (1980).

5

SINGULARITIES IN GENERAL RELATIVITY

In this chapter, we consider the question of nature and existence of space-time singularities in the general theory of relativity. After Einstein proposed the general theory describing the gravitational force in terms of space-time curvature and proposed the field equations relating the geometry and matter content of the space-time manifold, the earliest solutions found for the field equations were the Schwarzschild metric representing the gravitational field around an isolated body such as a spherically symmetric star, and the Friedmann cosmological models. Each of these solutions contained a space-time singularity where the curvatures and densities were infinite and the physical description would break down. In the Schwarzschild solution such a singularity was present at $r = 0$ whereas in the Friedmann models it was found at the epoch $t = 0$ which is beginning of the universe and origin of time where the scale factor $S(t)$ also vanishes and all objects are crushed to zero volume due to infinite gravitational tidal forces.

Even though the physical problem posed by the existence of such a strong curvature singularity was realized immediately in the solutions which turned out to have several important implications towards the experimental verification of the general relativity theory, initially this phenomena was not taken seriously. It was generally thought that the existence of such a singularity must be a consequence of the very high degree of symmetry imposed on the space-time while deriving these solutions. Subsequently, the distinction between a genuine singularity and a mere coordinate singularity became clear and it was realized that the singularity at $r = 2m$ in the Schwarzschild space-time was a coordinate singularity which could be removed by a suitable coordinate transformation. It was clear, however, that the genuine curvature singularity at $r = 0$ cannot be removed by any such coordinate transformation. The hope was then that when more general solutions are considered with a less degree of symmetry requirements, such singularities will be avoided.

This issue was sorted out when a detailed study of the structure of a general space-time and the associated problem of space-time singularity was taken up by Hawking, Penrose, and Geroch (see for example, Penrose, 1968; Geroch, 1971, and Hawking and Ellis, 1973). It was shown by this work that a space-time will admit singularities within a very general framework provided it satisfies certain reasonable assumptions such as the positivity of energy, a suitable causality assumption and a condition such as the existence of trapped surfaces. It thus follows that the space-time singularities form a general feature of the relativity theory. In fact, these considerations ensure the existence of singularities in other theories of gravity, which are based on a space-time manifold framework and satisfy the general conditions stated above.

In Section 5.1 we discuss first in some detail what is meant by a singular space-time and specify the notion of a singularity. It turns out that it is the notion of geodesic incompleteness that characterizes the notion of a singularity in an effective manner for a space-time and enables their existence to be proved by means of general theorems. A variety of ways in which a space-time exhibits singular behaviour, and the related notions of singular TIPs and TIFs (defined in Section 4.6), are discussed.

In Section 5.2 the gravitational focusing caused by the space-time curvature in congruences of timelike and null geodesics is discussed which turns out to be the main cause of the existence of singularity in the form of non-spacelike incomplete geodesics in space-time. The notions of expansion and shear for such a congruence and the related concept of conjugate points are introduced here. We then give an outline and the basic idea of the proofs for the theorems establishing the existence of space-time singularities in the form of geodesic incompleteness. We also describe here the further developments, especially in connection with the assumptions used in the singularity theorems. The singularity theorems establish the existence of the non-spacelike geodesic incompleteness for the space-time either in the past or future. However, they provide no information on the nature of these singularities or their properties. The open problems in this connection are outlined in this section, some of which will be further discussed in Chapters 6 and 7.

One of the important assumptions used by the singularity theorems is a suitable causality condition, such as the strong causality of the space-time. However, general relativity allows for situations where causality violation is permitted in a space-time. For example, the Gödel (1949) solution allows the existence of a closed timelike curve through every point of the space-time. One would of course like to rule out the causality violations on physical grounds, treating them as extremely pathological in that in such a case one would be able to enter one's own past. But the point is they are allowed in principle in general relativity. Then, can one avoid the space-time singularities if one allows for the violation of causality? This question is discussed in Section 5.3. We describe the results in this connection and point out that causality violation in its own right creates space-time singularities under certain conditions. Thus, this path of avoiding singularities does not look promising. We also show that for a causal space-time the violation of higher-order causality conditions gives rise to space-time singularities. Another question treated here is that of measure of causality violating sets when such a violation occurs. We show that in many cases, the causality violating sets in a space-time will have a zero measure, and thus such a causality violation may not be taken very seriously.

In Section 5.4 we consider the issue of physical nature of a space-time singularity. As pointed out in Section 5.2, there are many types of singular behaviours possible for a space-time and some of these could be regarded as mathematical pathologies in the space-time rather than having any physical significance. This will be especially so if the space-time curvature and similar other physical quantities remained finite along an incomplete non-spacelike geodesic in the limit of approach to the singularity. In this section we specify the criterion (Tipler, Clarke and Ellis, 1980) as to when a singularity should be considered physically important in terms of the curvature growth along singular geodesics. This criterion is used in Chapters 6 and 9 and the physical interpretation and implications of the same are considered.

5.1 Singular space-times

When should we say that a space-time manifold (M, g) is singular, or that M contains a space-time singularity? As pointed out in the beginning of this chapter, several examples of singular behaviour in the space-time models of general relativity are known. Important exact solutions of Einstein equations such as the Friedmann–Robertson–Walker cosmological models and the Schwarzschild space-time contain a space-time singularity where the energy density or the space-time curvatures diverge strongly and the usual description of the space-time breaks down.

In the Schwarzschild space-time there is an essential curvature singularity at $r = 0$ in the sense that along any non-spacelike trajectory falling into the singularity, as $r \to 0$, the Kretschman scalar $\alpha = R^{ijkl}R_{ijkl} \to \infty$. Also, all future directed non-spacelike geodesics which enter the horizon at $r = 2m$ must fall into this curvature singularity within a finite value of the proper time (finite value of the affine parameter in the case of null geodesics). Thus, all such curves are future geodesically incomplete.

In the Friedmann–Robertson–Walker models, the Einstein equations imply that if $\rho+3p > 0$ at all times, where ρ is the total energy density and p is the pressure, there is a singularity at $t = 0$ which could be identified as the origin of the universe. If $\rho + p > 0$ at all times then it is seen that along all the past directed trajectories meeting this singularity, $\rho \to \infty$ and also the curvature scalar $R = R_{ij}R^{ij} \to \infty$. Again, all the past directed non-spacelike geodesics are incomplete in the above sense. Thus, there is an essential curvature singularity at $t = 0$ which cannot be transformed away by any coordinate transformation. In fact, similar behaviour has been generalized to the class of spatially homogeneous cosmological models as shown by Ellis and King (1974) which satisfy the positivity of energy conditions $\rho \geq 0, \rho \geq 3p \geq 0$ and $1 \geq 3dp/d\rho \geq 0$.

The existence of such singularities where the curvature scalars and densities diverge imply a genuine space-time pathology where the usual laws of physics must break down. The existence of the geodesic incompleteness in these space-times imply that, for example, a timelike observer suddenly disappears from the space-time after a finite amount of proper time.

Of course, singular behaviour in space-times can also occur without bad behaviour of the curvature. A simple example is the Minkowski space-time with a point deleted. With such a hole in the space-time, there will be, for example, timelike geodesics running into the hole and hence be future incomplete. This is clearly an artificial situation which one would like to rule out in general by requiring that the space-time is *inextendible*, that is, cannot be isometrically embedded into another larger space-time manifold as a proper subset.

It is, however, possible to give a non-trivial example of the singular behaviour of above type where a conical singularity exists in the space-time as shown by Ellis and Schmidt (1977). Here the space-time is inextendible and the curvature components do not diverge in the limit of approach to the singularity. This is a behaviour similar to that occurring in a Weyl type of solution. The metric is given by

$$ds^2 = -dt^2 + dr^2 + r^2(d\theta^2 + \sin^2\theta\, d\phi^2), \tag{5.1}$$

with the range of coordinates given by $-\infty < t < \infty, 0 < r < \infty, 0 < \theta < \pi$ but with $0 < \phi < a$ with $\phi = 0$ and $\phi = a$ identified and $a \neq 2\pi$. There is a conical singularity at $r = 0$ through which the space-time cannot be extended and the singular boundary is related to the timelike two-plane $r = 0$ of the Minkowski space-time.

The important question that arises now is whether such singularities develop even when a space-time of generality is considered, and if so, under what conditions. In order to consider this question, it is first necessary to characterize more precisely what one means by a space-time singularity, or the singular behaviour for a general space-time, for the reasons to be stated below. Then it is seen that singularities must exist for a very wide class of space-times under a fairly reasonable general set of conditions. Such singularities may develop as the end state of gravitational collapse of a massive star, or in the cosmological situations such as the origin of the universe.

While trying to characterize a space-time singularity within the frame work of a space-time manifold of general relativity, the first point to note is that by very definition, the metric tensor has to be well defined at all the regular points of the space-time. Since this is no longer true at a space-time singularity such as those discussed above, a singularity cannot be regarded as a regular point of the space-time but must be treated as a boundary point attached to M. This situation causes difficulty when one tries to characterize a singularity by the criterion that the curvatures must blow up near the singularity. The trouble is, since the singularity is not a part of the space-time, it is not possible to define its neighbourhood in the usual sense to discuss the behaviour of curvature quantities in that region.

One might try to characterize the existence of a space-time singularity in terms of the divergence of the components of the Riemann curvature tensor along non-spacelike trajectories of the space-time. The trouble with this is that the behaviour of such components will, in general, change with the change of frames used and this approach is not really of much help. One could try using the curvature scalars or the scalar polynomials in the metric and the Riemann tensor and require them to achieve unboundedly large values. This is the behaviour encountered in the Schwarzschild and the Friedmann models. However, it is possible that such a divergence of curvature scalars occurs only at infinity for a given non-spacelike curve. In general it looks reasonable to demand that some sort of curvature divergence must take place along the non-spacelike curves which encounter a space-time singularity. We shall discuss this in more detail in Section 5.4. However, a general characterization of singularity in terms of the curvature divergence runs into various difficulties. For example, for the plane wave vacuum solutions the polynomials in curvature scalars vanish but the curvature tensor is still allowed to be singular (Penrose, 1965). Another example is the Taub-NUT type of solutions given by Misner (1963, 1967). Here the space-time curvatures are bounded, M is inextendible but it is both null and timelike geodesically incomplete.

Considering these as well as similar situations, the occurrence of non-spacelike geodesic incompleteness has been generally agreed upon as the criterion for the existence of a singularity for a space-time. This criterion does not cover all types of singular behaviours possible. For example, Geroch (1968a) has given an example of a space-time which is geodesically complete but contains a future inextendible timelike curve with a bounded acceleration and with finite proper length. This could correspond to a rocket ship with enough fuel to suddenly disappear from the universe after a finite proper time. Also, one would not like to term all the geodesically incomplete space-times as containing a physically genuine singularity, especially if the curvatures are everywhere finite throughout the space-time.

This includes the Taub–NUT case mentioned above. In order to call a singularity physically genuine, one would like to demand some sort of curvature divergence along the incomplete non-spacelike geodesic. On the other hand, if there is a powerful curvature divergence along an incomplete non-spacelike geodesic, one would certainly like to call such a singularity physically significant. We shall discuss this issue further in Section 5.4.

Nevertheless, it is clear that if a space-time manifold contains incomplete non-spacelike geodesics, there is a definite singular behaviour present in the space-time. In such a case, a timelike observer or a photon suddenly disappears from the space-time after a finite amount of proper time or after a finite value of the affine parameter. The singularity theorems which result from an analysis of gravitational focusing and global properties of a space-time prove this incompleteness property for a wide class of space-times under a set of rather general conditions.

There is a more general way in which one could define the singular points in a space-time using the terminal indecomposable pasts (TIPs) and terminal indecomposable futures (TIFs) defined in Section 4.6 (Penrose, 1974, 1979). The space-time is assumed to be strongly causal. By a curve we mean here a map γ from an interval $[0, a)$ of the real line into M where a could be possibly infinity. Thus, the curve starts at an initial point $\gamma(0)$ with a definite tangent but has no end point, as the interval is open at a. Such a curve will be called *extendible* if it is possible to extend the map γ to an end point $\gamma(a)$ in M and otherwise it is called *inextendible*. In particular, we are interested in the inextendible non-spacelike curves. The TIPs and TIFs are generated by future directed and past directed timelike curves respectively and they give all the boundary points of M which include both the singularities and points at infinity. Such a boundary point is called $\infty-$TIP, which is a point at infinity, if it is generated by some timelike curve of infinite proper time length in future. A *singular* TIP is the one which is not generated by any such timelike curve of infinite length (see Fig. 38). Similarly $\infty-$TIFS and singular TIFS can be defined. The existence of a singular TIP defines a singularity of space-time giving a class of future directed inextendible timelike curves which have finite proper time length but no future end point.

As pointed out by Clarke (1986), the basic requirement for the ideal end point of a timelike curve to be called a singularity, rather then a regular boundary point, is that there should be no extension of the space-time possible in which the curve in question could be continued. If such an extension existed, then the singularity would be similar in some sense to the coordinate singularity in the Schwarzschild geometry at $r = 2m$. Thus, the question of singularity depends on what type of extension is allowed for the space-time. Thus, a boundary point is called a C^k- *singularity* of

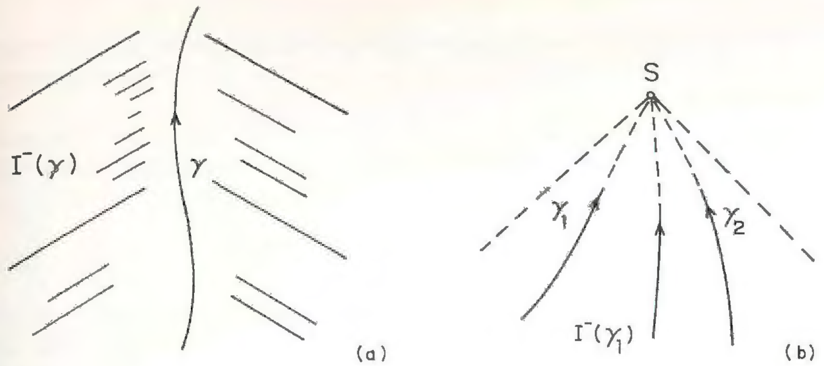

Fig. 38 The timelike curves γ and γ_1 in (a) and (b) respectively, are future inextendible without future end points. However, γ has an infinite length in future whereas γ_1 has a finite length. Thus in (a) the TIP $I^-(\gamma)$ represents a point at infinity and is a non-singular TIP, but in (b) the TIP $I^-(\gamma_1)$ is a singular TIP. Any other curve γ_2 defines the same singular TIP where we have $I^-(\gamma_1) = I^-(\gamma_2)$.

the space-time if there is no C^k extension of M which removes it. Clarke defines then the index k as a measure of strength of the singularity in the sense that the smaller the k is, the stronger is the singularity.

5.2 Gravitational focusing and singularities

Some important features of causal structure have been described in Chapter 4. We now describe how the matter fields with positive energy density affect the causality relations in a space-time and cause focusing in the families of timelike and null trajectories. The essential phenomena that occurs here is that matter focuses the non-spacelike geodesics of the space-time into pairs of focal points or the conjugate points. The basic property of conjugate points is that if $p < q$ are two conjugate points along a non-spacelike geodesic, then $p \ll q$. Now, there are null hypersurfaces such as the boundary of the future $I^+(p)$ for a point p such that no two points of such a hypersurface could be joined by a timelike curve. Thus, the null geodesic generators of such surfaces cannot contain conjugate points and must leave the hypersurface before encountering a conjugate point. This puts strong limits on such surfaces and the singularity theorems result from an analysis of such limits.

Consider now a congruence of timelike geodesics in the space-time. This is a family of curves such that through each point p there passes precisely one timelike geodesic trajectory. Choosing the curves to be smooth, this defines a smooth timelike vector field on the space-time. On the other hand, a given smooth vector field on the space-time specifies a congruence of curves in M.

Let V^i denote the timelike tangent vector to the congruence. Choosing the parameter to be the proper time along such timelike trajectories, this can be normalized to be a unit tangent vector, that is,

$$V^i V_i = -1. \tag{5.2}$$

The *spatial part* h_{ij} of the metric tensor can now be defined as

$$h_{ij} = g_{ij} + V_i V_j. \tag{5.3}$$

Then, $h^i{}_j = \delta^i{}_j + V^i V_j = g^{ik} h_{kj}$ and we can also see that

$$h_{ij} V^i = h_{ij} V^j = h^i{}_j V_i = h^i{}_j V^j = 0. \tag{5.4}$$

Thus, $h^i{}_j$ can be called the *projection operator* onto the subspace of T_p, orthogonal to the vector V^i. The indices of h are now raised and lowered just as in the case of the metric tensor,

$$h_{ij} h^j{}_k = (g_{ij} + V_i V_j)(g^j{}_k + V^j V_k) = g_{ik} + V_i V_k = h_{ik}, \tag{5.5}$$

and also we have

$$h^{ij} h_{ij} = h^i{}_i = \delta^i{}_i + V^i V_i = 3. \tag{5.6}$$

For the given congruence of timelike geodesics, the *expansion*, *shear*, and *rotation* tensors are respectively defined as below:

$$\theta_{ij} = V_{(k;l)} h_i{}^k h_j{}^l, \tag{5.7}$$

$$\sigma_{ij} = \theta_{ij} - \tfrac{1}{3} h_{ij} \theta, \tag{5.8}$$

$$\omega_{ij} = h_i{}^k h_j{}^l V_{[k;l]}. \tag{5.9}$$

Here, the *volume expansion* θ is defined by

$$\theta = \theta_{ij} h^{ij} = V_{(k;l)} h^{lk} = \nabla_k V^k = V^k{}_{;k}. \tag{5.10}$$

Further, we note that σ_{ij} and ω_{ij} are purely spatial quantities in the sense that

$$\sigma_{ij} V^i = \omega_{ij} V^i = 0.$$

Also, note that

$$\sigma^i{}_i = h^{ij} \sigma_{ij} = \theta - \tfrac{1}{3} h_{ij} h^{ij} = 0.$$

The covariant derivative of V is then expressed as

$$\nabla_j V_i = V_{i;j} = \tfrac{1}{3}\theta h_{ij} + \sigma_{ij} + \omega_{ij}, \tag{5.11}$$

This is verified by direct substitution from eqns (5.8), (5.9), and (5.10).

Now, the geodesic equation (2.52) implies,

$$V^k \nabla_k \nabla_j V_i = V^k \nabla_j \nabla_k V_i + R_{ilkj} V^l V^k.$$

Using the fact that V^k is the tangent to geodesics, that is, $\nabla_j (V^k \nabla_k V_i) = 0$, the above equation can be written as

$$V^k \nabla_k \nabla_j V_i = -(\nabla_j V^k)(\nabla_k V_i) + R_{ilkj} V^l V^k.$$

Taking trace in the above we get

$$\frac{d\theta}{d\tau} = V^k \nabla_k V^i{}_{;i} = -(V^k{}_{;i} V^i{}_{;k}) - R_{lk} V^l V^k \tag{5.12}$$

Using eqn (5.11) in the above and the anti-symmetry properties of the tensor ω_{ij} we get after some simplification

$$\frac{d\theta}{d\tau} = -R_{lk} V^l V^k - \tfrac{1}{3}\theta^2 - \sigma_{ij}\sigma^{ij} + \omega_{ij}\omega^{ij}$$

which can be written as

$$\frac{d\theta}{d\tau} = -R_{lk} V^l V^k - \tfrac{1}{3}\theta^2 - 2\sigma^2 + 2\omega^2. \tag{5.13}$$

Equation (5.13) is called the *Raychaudhuri equation* (Raychaudhuri, 1955) which describes the rate of change of the volume expansion as one moves along the timelike geodesic curves in the congruence.

The second and third term on the right-hand side involving θ and σ are positive always. Consider now the term $R_{ij} V^i V^j$. By Einstein equations this can be written as

$$R_{ij} V^i V^j = 8\pi [T_{ij} V^i V^j + \tfrac{1}{2} T]. \tag{5.14}$$

The term $T_{ij} V^i V^j$ above represents the energy density as measured by a timelike observer with the unit tangent V^i, which is the four-velocity of the observer. For all reasonable classical physical fields this energy density is generally taken as non-negative and it is assumed that for all timelike vectors V^i the following is satisfied

$$T_{ij} V^i V^j \geq 0. \tag{5.15}$$

Such an assumption is called the *weak energy condition*. On the other hand, it is also considered reasonable to believe that the matter stresses will not be so large as to make the right-hand side of eqn (5.14) negative. This will be satisfied when the following is satisfied,

$$T_{ij}V^iV^j \geq -\tfrac{1}{2}T. \tag{5.16}$$

Such an assumption is called the *strong energy condition* and it implies that for all timelike vectors V^i,

$$R_{ij}V^iV^j \geq 0. \tag{5.17}$$

By continuity it can be argued that the same will then hold for all null vectors as well.

Both the strong and weak energy conditions will be valid for well-known forms of matter such as the perfect fluid provided the energy density ρ is non-negative and there are no large negative pressures which are bigger or comparable to ρ (when converted into physical units; see for example, Wald (1984), for further discussion.)

An additional energy condition required often by the singularity theorems is the *dominant energy condition*, which states that in addition to the weak energy condition, the pressure of the medium must not exceed the energy density. This is equivalently stated as, for all timelike vectors V^i, $T_{ij}V^iV^j \geq 0$ and the vector $T^{ij}V_i$ is a non-spacelike vector. Such a condition would be satisfied provided the local speed of sound does not exceed the local speed of light.

With the strong energy condition being satisfied, the Raychaudhuri equation implies that the effect of matter on space-time curvature causes a focusing effect in the congruence of timelike geodesics due to gravitational attraction. This, in general causes the neighbouring geodesics in the congruence to cross each other to give rise to caustics or conjugate points. This separation between nearby timelike geodesics is governed by the geodesic deviation equation,

$$D^2Z^j = -R^j{}_{kil}V^kZ^iV^l$$

where Z^i is the separation vector between nearby geodesics of the congruence. Solutions of the above equation are called the *Jacobi fields* along a given timelike geodesic.

Suppose now γ is a timelike geodesic. Then, two points p and q along γ are called *conjugate points* if there exists a Jacobi field along γ which is not identically zero but vanishes at p and q. From the derivation of the Raychaudhuri equation given above it is clear that the occurrence of conjugate points along a timelike geodesic is closely related to the behaviour

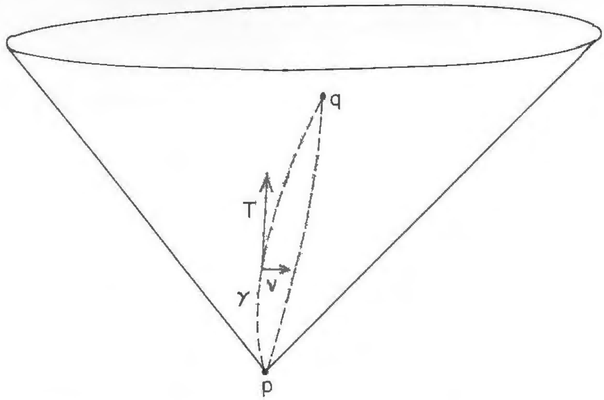

Fig. 39 Infinitesimally separated null geodesics cross at p and q, which are conjugate points along the curve γ.

of the expansion parameter θ of the congruence. In fact, it can be shown that the necessary and sufficient condition for a point q to be conjugate to p is that for the congruence of timelike geodesics emerging from p, we must have $\theta \to -\infty$ at q (see for example, Hawking and Ellis, 1973). The conjugate points along null geodesics are also similarly defined. Consider for example, a congruence of null geodesics emanating from a point p. If infinitesimally near by null geodesics of the congruence meet again at some other point q in future, then p and q are called *conjugate* to each other (Fig. 39).

Similarly, let S be a smooth spacelike hypersurface in M, that is, it is an embedded three-dimensional submanifold. Consider a congruence of timelike geodesics orthogonal to S. Then a point p along a timelike geodesic γ of the congruence is called *conjugate to* S along γ if there exists a Jacobi vector field along γ which is non-zero at S but vanishes at p. This means that there are two infinitesimally nearby geodesics orthogonal to S which intersect at p (Fig. 40). Again, the equivalent condition for this to happen, in terms of the parameter θ is that the expansion θ for the congruence orthogonal to S tends to $-\infty$ at p. If V^i denotes the unit timelike tangent vector field of the congruence of timelike geodesics, where V^i denotes the normal to S, then the *extrinsic curvature* χ_{ij} of S is defined as

$$\chi_{ij} = \nabla_i V_j, \tag{5.18}$$

which is evaluated at S. Clearly, $\chi_{ij}V^i = \chi_{ij}V^j = 0$. Also, the hypersurface orthogonality of the congruence implies that $\omega_{ij} = 0$. As a result, $\chi_{ij} = \chi_{ji}$, that is, this is a symmetric tensor. The trace of the extrinsic

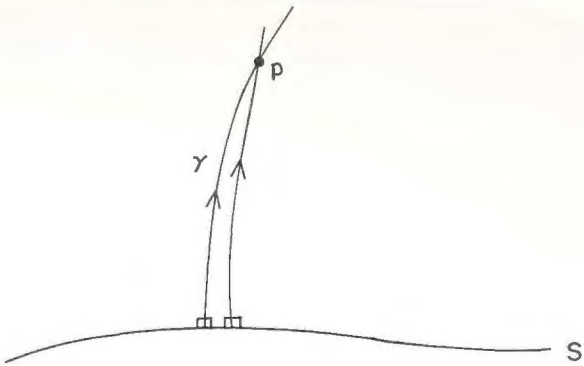

Fig. 40 A point p conjugate to the spacelike hypersurface S. The timelike geodesic γ is orthogonal to S, which is intersected by another infinitesimally near by timelike geodesic.

curvature, denoted by χ is given by

$$\chi = \chi^i{}_i = h^{ij}\chi_{ij} = \theta. \tag{5.19}$$

Thus, we have $\chi = \theta$ at S where θ is the expansion of the congruence orthogonal to S.

Now, the behaviour of the expansion parameter θ is governed by the Raychaudhuri equation as pointed out above. For example, consider the situation when the space-time satisfies the strong energy condition and the congruence of timelike geodesics is hypersurface orthogonal. In such a case $\omega_{ij} = 0$ and the corresponding term ω^2 vanishes in eqn (5.13). Then, the expression for the covariant derivative of ω_{ij} implies that it must vanish for all future times as well. It follows from the above discussion that we must have

$$\frac{d\theta}{d\tau} \leq -\frac{\theta^2}{3},$$

which means that the volume expansion parameter must be necessarily decreasing along the timelike geodesics. If θ_0 denotes the initial value of the expansion then the above can be integrated as $\theta^{-1} \geq \theta_0^{-1} + \tau/3$. It is clear from this that if the congruence is initially converging and θ_0 is negative then $\theta \to -\infty$ within a proper time distance $\tau \leq 3/\mid \theta_0 \mid$.

The following result then follows from the above discussion.

Proposition 5.1. Let M be a space-time satisfying the strong energy condition and S be a spacelike hypersurface with $\theta < 0$ at $p \in S$. If γ is the timelike geodesic of the congruence orthogonal to S passing through p, then there exists a point q conjugate to S along γ within a proper time distance $\tau \leq 3/\mid \theta \mid$, provided γ can be extended to that value of the proper time.

Suppose now that the trace of extrinsic curvature χ_{ij} (which is also some times called the *second fundamental form* of the surface S) is everywhere negative on S, that is, $\theta = \chi < 0$ on S, and further, it is bounded above by a negative value θ_{max}. In that case, it is clear from the above discussion that all the timelike geodesics of the congruence orthogonal to S will contain a point conjugate to S within a proper time distance $\tau \leq 3/\mid \theta_{max} \mid$. Thus, we have the following.

Proposition 5.2. Let M be a space-time satisfying the conditions of Proposition 5.1 and S be a spacelike surface in M. Let the trace of the extrinsic curvature $\chi = \theta < 0$ on S and bounded above by a negative value θ_{max}. Then all the timelike geodesics orthogonal to S have a point p conjugate to S within a proper time distance $\tau \leq 3/\mid \theta_{max} \mid$, provided the geodesics can be extended to that value of the proper time.

Consider now the congruence of timelike geodesics passing through a space-time point p. As shown by Lemma 4.5.2 of Hawking and Ellis (1973), for any convex normal neighbourhood of p, the trajectories of this congruence are orthogonal to the spacelike surfaces of proper time $\tau =$ const. along the geodesics. Thus, the congruence is hypersurface orthogonal and $\omega_{ij} = 0$, and it will be zero for all future times as well. Then the above discussion again implies the following result for the occurrence of conjugate points along a timelike geodesic γ.

Proposition 5.3. Let M be a space-time satisfying the strong energy condition for all timelike vectors V^i. Let $p \in \gamma$ for a timelike geodesic γ and suppose $\theta = \theta_0 < 0$ at some point q in future of p along γ. Then γ contains a point r conjugate to p within a proper time distance $\tau \leq 3/\mid \theta_0 \mid$ from q, provided it can be extended to that value of proper time.

The basic implication of the above results is that once a convergence occurs in a congruence of timelike geodesics, the conjugate points or the caustics must develop in the space-time. These can be interpreted as the singularities of the congruence. Such singularities could occur even in Minkowski space-time and similar other perfectly regular space-times. However, when combined with certain causal structure properties of space-time as stated in Chapter 4, the results above imply the existence of space-time singularities in the form of geodesic incompleteness.

One could similarly discuss the gravitational focusing effect for the congruence of null geodesics or for null geodesics orthogonal to a spacelike two-surface. In this case, it is instructive to use the coordinate u introduced in the previous chapter so that the surfaces $u =$ const. describe non-intersecting null hypersurfaces in the space-time. Then, the normal vector to these hypersurfaces is given by

$$l_i = \partial_i u = (1, 0, 0, 0),$$

which is a null vector as we have

$$l^i l_i = 0.$$

Then, the equation for null geodesics can be written as

$$l^i{}_{;j} l^j = 0.$$

One can now again define the *expansion* θ, *rotation* ω, and the *shear* σ for the null congruence as

$$\theta = \tfrac{1}{2} l^i{}_{;i}, \quad \omega = [\tfrac{1}{2} l_{[i;j]} l^{i;j}]^{1/2},$$

and

$$\mid \sigma \mid = [\tfrac{1}{2} l_{(i;j)} l^{i;j} - \theta^2]^{1/2}.$$

It is again possible to write down a null version of the Raychaudhuri equation using the above quantities and one could discuss the occurrence of conjugate points along these trajectories. For further details we refer to Hawking and Ellis (1973) and Wald (1984). It is interesting to note that in the case of a null geodesic congruence, the quantities θ, ω, and σ defined above are called *optical scalars* because of their following physical interpretation due to R. K. Sachs. Consider an object placed in the null congruence, and a screen at an infinitesimal distance dr from the object. Then, the shape, size, and orientation of the shadow of the object depends only on the location of the screen and not its velocity, and the shadow is expanded by θdr, twisted by $\mid \sigma \mid dr$ and rotated by ωdr. We now note the relevant result in this connection

Proposition 5.4. Let M be a space-time satisfying $R_{ij}K^iK^j \geq 0$ for all null vectors K^i and γ be a null geodesic of the congruence. If the convergence θ of null geodesics from some point p is $\theta = \theta_0 < 0$ at some point q along γ, then within an affine distance less than or equal to $2/ \mid \theta_0 \mid$ from q the null geodesic γ will contain a point conjugate to p, provided it can be extended to that affine distance.

In fact, the above result on the occurrence of conjugate points could be strengthened further by observing that even if $R_{ij}V^iV^j = 0$ throughout the space-time, if $\sigma^2 > 0$ then a net focusing effect again results. This will be so if $R_{ijkl}V^jV^l \neq 0$ at least at one point in the space-time. It is then possible to show the following result (for a proof, we refer to Hawking and Ellis, 1973).

Proposition 5.5. Let $\lambda(t)$ be a non-spacelike geodesic which is complete both in future as well as in the past, that is, the affine parameter t varies over $(-\infty, +\infty)$. Let $R_{ij}V^iV^j + 2\sigma^2$ be continuous and non-negative on $(-\infty, +\infty)$. If $R_{ij}V^iV^j + 2\sigma^2 > 0$ for at least one value of t , then $\lambda(t)$ must contain a pair of conjugate points in the interval $(-\infty, +\infty)$.

The condition required in the above proposition essentially amounts to the statement that the non-spacelike trajectory must pass through some matter or radiation at least once throughout its history. This will happen if $R_{ij}V^iV^j \neq 0$ at some point on the trajectory or the Weyl tensor is non-zero at this point in a suitable manner. A precise condition to ensure this is that every non-spacelike geodesic in M must contain a point at which $K_{[i}R_{j]el[m}K_{n]}K^eK^l \neq 0$, where K is the tangent to the non-spacelike geodesic. This is called the *generic condition* on M. Thus every timelike and null geodesic in M, which is both future and past complete, must contain a pair of conjugate points if the space-time satisfies the generic condition.

We have discussed globally hyperbolic space-times in Chapter 5 and an important property of these space-times pointed out there is if N is a globally hyperbolic subset of M and if $p \ll q$ for $p, q \in N$ then there exists a timelike geodesic from p to q which maximizes the lengths of all non-spacelike curves from p to q. Such a maximal curve is related to the existence of conjugate points in the following manner (Hawking and Ellis, 1973).

Proposition 5.6. Let N be a globally hyperbolic subset of M and $p, q \in N$. Let $\gamma(t)$ be a timelike geodesic which maximizes the lengths of all non-spacelike curves from p to q. Then $\gamma(t)$ contains no points conjugate to p between p and q.

The point is that, if there were a conjugate point r to p between p and q, then it can be shown using variational arguments that one could get a longer non-spacelike curve from p to q by 'rounding off the corner' at r generated due to the conjugate point r. This is contradictory to the maximality of the curve $\gamma(t)$.

There are several singularity theorems available which establish the non-spacelike geodesic incompleteness for a space-time under different sets of conditions and applicable to different physical situations. However, the most general of these is the Hawking–Penrose theorem (Hawking and Penrose, 1970), which is applicable in both the collapse situation and cosmological scenario. The main idea of the proof of such a theorem is the following. Using the causal structure analysis it is shown that there must be maximal length timelike curves between certain pairs of events in the space-time. Now, as pointed out above, a causal geodesic which is both future and past complete must contain pairs of conjugate points if M satisfies the generic condition and an energy condition. This is then used to draw the necessary contradiction to show that M must be non-spacelike geodesically incomplete.

Theorem 5.7 (Hawking and Penrose, 1970). A space-time (M, g) cannot be timelike and null geodesically complete if the following are satisfied:

(1) $R_{ij}K^iK^j \geq 0$ for all non-spacelike vectors K^i;
(2) the generic condition is satisfied, that is, every non-spacelike geodesic contains a point at which $K_{[i}R_{j]el[m}K_{n]}K^eK^l \neq 0$, where K is the tangent to the non-spacelike geodesic;
(3) the chronology condition holds on M; and
(4) there exists in M either a compact achronal set without edge or a closed trapped surface, or a point p such that for all past directed null geodesics from p, eventually θ must be negative.

The main idea of the proof is the following. One shows that the following three cannot hold simultaneously:

(a) every inextendible non-spacelike geodesic contains pairs of conjugate points;
(b) the chronology condition holds on M;
(c) there exists an achronal set S in M such that $E^+(S)$ or $E^-(S)$ is compact.

If this is shown then the theorem is proved because (3) is same as (b), (4) implies (c), and (1) and (2) imply (a). First, we note that (a) and (b) imply the strong causality of the space-time (see proposition 6.4.6 of Hawking and Ellis, 1973). Next, it can be shown that if S is a future trapped set and if strong causality holds on $\overline{I^+(S)}$ then there exists a future endless trip γ such that $\gamma \subset IntD^+(E^+(\gamma))$. Now, one defines $T = \overline{J^-(\gamma)} \cap E^+(S)$, then T turns out to be past trapped and hence there exists λ, a past endless causal geodesic in $Int(D^-(E^-(T))$. Then one chooses a sequence $\{a_i\}$ receding into the past on λ and a sequence $\{c_i\}$ on γ to the future. The sets $J^-(c_i) \cap J^+(a_i)$ are compact and globally hyperbolic, so there exists a maximal geodesic μ_i from a_i to c_i for each i. The intersections of μ_i with the compact set T have a limit point p and a limiting causal direction. The causal geodesic μ with this direction at p must have a pair of conjugate points. This is then shown to be contradictory to the maximality property of the geodesics stated above.

5.3 Singularities and causality violation

The singularity theorem stated above and other singularity theorems as well contain the assumption of causality or strong causality. This offers the alternative that causality may be violated rather than a singularity occurring in the space-time. Hence, the implication of the singularity theorem stated above is that when there is enough matter present in the universe either the causality must be violated or a boundary point must exist for the space-time. In the cosmological case, such stress-energy density will be provided by the microwave background radiation (Hawking and Ellis, 1973) or in the case of stellar collapse trapped surfaces may form (Schoen and Yau, 1983) providing a condition leading to the formation of a singularity. The Einstein equations as such do not rule out the causality violating configurations which depend on the global topology of the space-time. Hence the question of causality violations versus space-time singularity needs a careful examination.

This question was examined for finite causality violations in asymptotically flat space-times by Tipler (1976, 1977a), who showed that in this case the causality violation in the form of closed timelike lines is necessarily accompanied by incomplete null geodesics, that is, a space-time singularity, provided the strong energy condition is satisfied for all null vectors and the generic condition is satisfied. He also assumed further that the energy density ρ has a minimum strictly greater than zero along past directed null geodesics. Further results on causality violations and higher-order causality violations with reference to occurrence of singularities have been reported by Joshi (1981), and Joshi and Saraykar (1986), who show that the causality violations must be accompanied by singularities even when the space-time is causal but the higher-order causality conditions are violated. Clarke and Joshi (1988) studied global causality violation for a reflecting space-time and the theorems of Kriele (1990) have improved some of the conditions under which the results on chronology violations implying the singularities have been obtained. Also, global causality violating space-times have been studied by Clarke and de Felice (1982).

Taken as a whole, the above results imply that violating either causality or any of the higher-order causality conditions may not be considered a good alternative to the occurrence of space-time singularities. There are philosophical problems connected with the issue of causality violation such as entering one's own past when the causality is violated. However, even if one allowed for the causality violations, the above results show that these are necessarily accompanied by singularities of their own.

In the following, we study the higher-order causality violations in some detail and also the issue of measure of a causality violating set in a space-time. Even though sometimes arguments are proposed to favour the causality violation over the occurrence of a space-time singularity, the results of Tipler can be interpreted to show the physical unreasonableness of causality violations in the sense of inducing the space-time singularities in the form of incomplete null geodesics. However, even if closed timelike curves did not exist in a space-time, it would not mean that the space-time is causally well-behaved. As pointed out in Chapter 4, there is a hierarchy of higher-order causality conditions for a space-time such as the strong causality, stable causality, causal continuity and so on, and violation of any one of them may be regarded as a serious causal pathology. Such a space-time could be on the verge of violating the causality, for example, in the case of stable causality violation, and hence it would be interesting to examine the status of a space-time in which higher-order causality violations occur but which is otherwise causal.

We assume the matter tensor to satisfy the following condition (which is similar to the conditions discussed by Tipler (1977a) on the null geodesics of the space-time. Let $\gamma(s)$ be any past directed null geodesic in M. Then

$$\lim_{s\to k}\ \inf\ T_{ij}K^iK^j > 0, \tag{5.20}$$

must hold along γ where k is the limit of the affine parameter in the past. Such a condition may be considered reasonable in the cosmological setting when the matter and radiation are present, for example, the microwave background radiation, which should have higher densities in the past in view of the observed expansion of the universe. In fact, one could weaken the above condition by using the results on occurrence of conjugate points on half-geodesics (see for example, Borde, 1987) or using the methods such as those used by Kriele (1990). However, we shall not consider such a generalization here and use the condition (5.20) above for the simplicity of presentation and because it conveys the basic idea of such results.

Suppose now M is a causal space-time which satisfies the Einstein equations. Since we do not assume M to be asymptotically flat, the result is relevant to the cosmological case as well. Let one of the higher-order causality conditions such as the strong causality, stable causality, causal continuity, or causal simplicity be violated in M. Then M cannot be causally simple as that would imply the validity of all the other higher-order causality conditions stated above as discussed in Chapter 4. So there exists a closed set $S \subset M$ on which the causal simplicity is violated, that is, $J^+(S)$ is not closed. Clearly, S must be a proper subset of M, because if $S = M$ then $S = J^+(S) = M$ which is closed. Then, the boundary $\dot{I}^+(S)$ must be non-empty in M because, $J^+(S)$ not being closed, $\overline{J^+(S)} \neq M$ and it must be a proper subset of M. Then, $\overline{J^+(S)} \subseteq \overline{M} = M$ and there exists an event $p \in \overline{J^+(S)}$ which is not in $J^+(S)$. This implies that

$p \in \dot{J}^+(S) = \dot{I}^+(S)$. We note that this will not be the case necessarily when causality is violated in M; for example, in the case of the Gödel universe, $I^+(S) = M$ as the causality violation is total, in which case $\dot{I}^+(S) = \emptyset$. It follows that there is a point $q \in \dot{I}^+(S)$ such that no null geodesic from q meets S. Then, the null generator γ of $\dot{I}^+(S)$ with the future end point at q must be past endless in $\dot{I}^+(S)$ and does not meet S.

Suppose now that the generator γ is past complete and we assume the weak energy condition, that is, $T_{ij}W^iW^j \geq 0$ for all non-spacelike vectors W^i. Then we have $R_{ij}K^iK^j \geq 0$ for all null vectors K^i. Referring now to the Raychaudhuri equation (5.13), since the quantity σ^2 is intrinsically positive, eqn (5.20) implies that along γ

$$\int_0^\infty \tfrac{1}{2}(R_{ij}K^iK^j + 2\sigma^2)ds = \infty. \tag{5.21}$$

However, in this situation the null trajectory γ must contain infinitely many conjugate points (Tipler, 1977a) and any two of such points can be timelike related. This is contradictory to the achronal nature of the null hypersurface $\dot{I}^+(S)$. Hence γ must be incomplete in the past.

Of all the higher-order causality conditions stated above, much physical importance can be attached to the stable causality condition which ensures that if M is causal, its causality property should not be disturbed with small perturbations in the metric tensor. Presumably, the general theory of relativity is a classical approximation to some, as yet unknown, quantum theory of gravity in which the value of the metric at a point will not be exactly known and small fluctuations in the value must be taken into account. Thus we have seen that if the causality of M breaks down with the slight perturbation of the metric then this must be accompanied by the occurrence of space-time singularities.

A space-time M is called global chronology violating or totally vicious if there is an event $p \in M$ such that $I^+(p) = I^-(p) = M$. One could see that the above argument will also apply, when M is not global chronology violating, to show that closed timelike lines are necessarily accompanied by the occurrence of singularities without requiring the assumption of asymptotic flatness. Such a result would be relevant for the cosmological situation. Suppose chronology is violated at some $p \in M$ and γ is the closed timelike curve through p. If now $q \in J^+(p)$ then there is a non-spacelike curve λ from p to q. Combining γ and λ we can get a timelike curve from p to q as pointed out in Section 4.1 and hence $q \in I^+(p)$. It follows that $J^+(p) = I^+(p)$, $J^+(p)$ is an open set in M, and $\dot{I}^+(p) = \dot{J}^+(p)$. Since M is not global chronology violating, $I^+(p) = J^+(p) \neq M$ and $J^+(p)$ is a proper subset of M. Hence, its closure is contained in M and there is an event $r \in \overline{J^+(p)}$ which is not in $J^+(p)$. Thus $r \in \dot{I}^+(p)$. In that case, there is a generator of this boundary with r as the future endpoint which cannot meet p in the past because that implies $r \in J^+(p)$. Thus, this generator is past endless in $\dot{I}^+(p)$. Then, an argument similar to the above shows that this generator must be past incomplete provided the weak energy condition and eqn (5.20) are satisfied. It follows that chronology violations in a space-time give rise to singularities in rather general situations.

Independently of the relationship between singularities and causality violations, another important question is the measure of causality violating sets in a space-time. The point is, even if the causality is violated in a space-time, if such a violation occurs in an isolated manner only on a set of measure zero, it may not be considered physically significant. However, it is known that whenever chronology is violated, that will happen for a set of non-zero measure. If a closed timelike curve passed through $p \in M$ then $p \in I^+(p)$ and $I^+(p)$ being an open set, it contains an open neighbourhood N of p such that $N \subset I^+(p)$. Then, by varying the given closed timelike curve one could create causality violation in all of N.

In the following, we examine in some detail the question of the nature of a set $D \subset M$ on which some of the higher-order causality conditions are violated. For example, the distinguishingness of the space-time is a very useful condition in that it is crucial to many global constructions such as the space-time boundary as given by Geroch, Kronheimer and Penrose (1972), and taken together with reflectingness it gives the important criterion of causal continuity. This condition not only rules out closed timelike curves but also prohibits almost closed non-spacelike curves. Following Joshi (1985) and Vyas and Akolia (1986), we study some properties of the set $\triangle \subset M$ where distinguishingness fails. Here $\triangle = \triangle^+ \cup \triangle^-$, where $\triangle^+$ and $\triangle^-$ are the sets on which future distinguishingness and past distinguishingness are violated respectively in M.

For $x \in M$ we define the sets

$$\triangle^+(x) = \{y : I^+(x) = I^+(y); y \neq x\},$$

$$\triangle^-(x) = \{y : I^-(x) = I^-(y); y \neq x\}.$$

It is seen that $\triangle^+$ is a disjoint union of the sets $\triangle^+(x)$ and similar result holds for $\triangle^-$. It is clear from the definition that $\triangle^+ = \cup \triangle^+(x)$. Suppose $p \in \triangle^+(x) \cap \triangle^+(y)$. Then, by definition, $I^+(x) = I^+(p)$ and $I^+(y) = I^+(p)$, which implies $I^+(x) = I^+(y)$. It follows that for any event p, $p \in \triangle^+(x)$ if and only if $p \in \triangle^+(y)$, which shows that $\triangle^+(x) = \triangle^+(y)$. It follows that the sets $\triangle^+(x)$ give rise to a partition of $\triangle^+$.

Just as the chronology violating set is necessarily open, it can be shown (Vyas and Akolia, 1986) that the sets $\triangle^+(x)$ are closed provided M satisfies the following symmetry condition: $x \in \triangle^+(x)$ if and only if $x \in \triangle^-(x)$.

Of course, the closure of $\Delta^{\pm}$ does not follow from this because an arbitrary union of closed sets need not be closed. However, these sets will be closed if M is reflecting, because as shown by Penrose (1972), the region on which strong causality is violated is closed and as shown in Chapter 4, the distinguishingness and strong causality conditions are equivalent for a reflecting space-time.

Examining the nature of Δ^+, it is seen that no point of this set will be an isolated point in the space-time. Suppose $p \in \Delta^+$. Then there exists $x \in M$ such that $I^+(p) = I^+(x)$, with $p \neq x$. We assume the space-time to be causal. Then p cannot be chronologically related to x. However, $I^+(p) \subseteq I^+(x)$ implies $p \in \dot{I}^+(x)$. Then, as discussed in Chapter 4, there is a null generator γ of $\dot{I}^+(x)$ with a future end point at p. Let $t \in \gamma$. Then $t < p$ and $t \in \dot{I}^+(x)$. Therefore, $I^+(p) \subset I^+(t)$ and $I^+(t) \subset I^+(x)$ with t not in $I^+(x)$. But $I^+(p) = I^+(x)$ implies that $I^+(p) = I^+(t)$. Thus $t \in \Delta^+(p)$. It follows that $\gamma \subset \Delta^+$.

If one could show that the interior of the set Δ^+ is empty then it would follow that, with a suitable definition of measure, the region of space-time on which the distinguishingness is violated is a zero measure set. This can be achieved if we assume the non-imprisonment condition on the space-time. Suppose the interior of Δ^+ is non-empty and p belongs to this interior set. Then there is a neighbourhood N of p such that $N \subset \Delta^+$. Hence there is a future directed non-spacelike curve λ from p which leaves N and re-enters it. If $x \in \lambda \cap N \subset \Delta^+$ then again there is a non-spacelike curve λ_1 from x which leaves and re-enters N. Continuing this process gives a curve which leaves and re-enters N for an infinite number of times, which violates the non-imprisonment condition on M. In fact, if we assume the causality of M, we could show that the interior of $\Delta^+(x)$ must be empty for all $x \in M$. This will follow if we show that no point of $I^+(x)$ is in $\Delta^+(x)$ because every neighbourhood of x must contain points of $I^+(x)$. Let $p \in I^+(x)$ such that $p \in \Delta^+(x)$ as well. Then, by definition of $\Delta^+(x)$, $I^+(x) = I^+(p)$ and $p \in I^+(x)$ implies $p \in I^+(p)$, which is the violation of causality.

The above discussion shows when the distinguishingness violating set of the space-time will have empty interior. One could also consider the set on which the reflecting condition is violated, and similar to the violation of the distinguishingness, it can be shown that if R denotes the set of points in M at which the reflecting condition is violated, then no point of R is an isolated point in M (Vyas and Akolia, 1986). In fact, Ishikawa (1979) pointed out that the region on which the reflecting condition is violated is a subset of M with empty interior, however, it could possibly be dense in some open subset of M.

5.4 Strong curvature singularities

As pointed out in Section 5.1, the existence of an incomplete non-spacelike geodesic, or the existence of an inextendible non-spacelike curve which has a finite length as measured by a generalized affine parameter, implies the existence of a space-time singularity. The *generalized affine length* for such a curve is defined as (Hawking and Ellis, 1973),

$$L(\lambda) = \int_0^a \left[\sum_{i=0}^{3}(X^i)^2\right]^{1/2} ds, \tag{5.22}$$

which is a finite quantity. The X^is are the components of the tangent vector to the curve in a parallely propagated tetrad frame along the curve. Each such incomplete curve defines a boundary point of the space-time which is a singularity. In order to call this a genuine physical singularity, one would typically like to associate such a singularity with unboundedly growing space-time curvatures. If all the curvature components and the scalar polynomials formed out of the metric and the Riemann curvature tensor remained finite and well-behaved in the limit of approach to the singularity along an incomplete non-spacelike curve, it may be possible to remove such a singularity by extending the space-time when the differentiability requirements are lowered (Clarke, 1986).

There are several ways in which such a requirement can be formalized. For example, a *parallely propagated curvature singularity* is the one which is the end point of at least one non-spacelike curve on which the components of the Riemann curvature tensor are unbounded in a parallely propagated frame. On the other hand, a *scalar polynomial singularity* is the one for which a scalar polynomial in the metric and the Riemann tensor takes an unboundedly large value along at least one non-spacelike curve which has the singular end point. This includes the cases such as the Schwarzschild singularity where the Kretschman scalar $R^{ijkl}R_{ijkl}$ blows up in the limit as $r \to 0$.

What is the guarantee that such curvature singularities will at all occur in general relativity? The answer to this question, for the case of parallely propagated curvature singularities, is provided by a theorem of Clarke (1975) which establishes that for a globally hyperbolic space-time M which is C^{0-} inextendible, when the Riemann tensor is not very specialized in the sense of not being type-D and electrovac at the singular end point, then the singularity must be a parallely propagated curvature singularity. For another class of such curvature singularities to be characterized below, we shall show in Chapter 6 that they arise for a wide range of space-times involving gravitational collapse.

A class of physically relevant singularities, called the strong curvature singularities was defined and analysed by Tipler (1977b); Tipler, Clarke and Ellis (1980), and Clarke and Królak (1986). The idea here is to define a physically all embracing strong curvature singularity in such a way so that all the objects falling within the singularity are destroyed and crushed to zero volume by the infinite gravitational tidal forces. This notion can be formulated in the following manner: let $\lambda(t)$ be a timelike or null geodesic which is incomplete at an affine parameter value $t = 0$. Let K^i denote the tangent vector to λ and $\mu(t) = Z_1 \wedge Z_2 \wedge Z_3$ be a volume form defined along $\lambda(t)$ where Z_1, Z_2, Z_3 are linearly independent Jacobi vectors defined along the curve λ orthogonal to K^i. (If λ is null then $\mu(t)$ is defined as a two-form.) A real valued map from the space of all such three-forms can be defined by $\Delta(A \wedge B \wedge C) = \det[A^i, B^i, C^i]$. We denote $\Delta(\mu(t))$ by $V(t)$, which defines a volume element along $\lambda(t)$ and is independent of choice of basis. The singularity at $t = 0$ is then called a *strong curvature singularity* if $V(t) = 0$ in the limit as $t \to 0$ for all possible $\mu(t)$, that is, for all possible choices of linearly independent Jacobi fields.

This definition effectively captures the notion that all objects falling into a strong curvature singularity are crushed to zero volume. A singularity will be called a strong curvature singularity provided there exists at least one non-spacelike geodesic terminating at the singularity along which the above curvature condition is satisfied. However, there is a criterion whereby a strong curvature singularity is defined in a much stronger sense (Tipler, 1977b) by requiring that the strong curvature condition above must be satisfied along *all* the non-spacelike geodesics terminating into the singularity. Specifically, necessary and sufficient conditions for the occurrence of strong curvature singularities are derived by Clarke and Królak (1986), which are shown to involve the tetrad components of Riemann, Ricci, and Weyl tensors and also the divergence of their integrals along non-spacelike geodesics running into the singularity. It follows from their analysis that an incomplete non-spacelike geodesic does not define a strong curvature singularity unless either the Weyl or the Ricci tensor components diverge sufficiently fast along such a trajectory.

A sufficient condition for the same to happen is that, if t is the affine parameter, we must have

$$R_{ij} V^i V^j \geq A/t^2, \tag{5.23}$$

for some fixed constant A along the trajectory in the limit of approach to the singularity as $t \to 0$. This provides a sufficient condition for all the 2-forms $\mu(k)$ defined along the singular null geodesic to vanish as singularity is approached and implies a very powerful curvature growth establishing a strong curvature singularity. For the timelike geodesics this will imply that all the volume forms defined by the Jacobi fields along these trajectories must vanish in the limit of approach to the singularity, or they must vanish infinitely many times in this limit.

To fix the ideas, we consider the case $R_{ij}V^iV^j = K/t^2$ in some detail. It is convenient to define a length scale y associated with the volume $V(t)$ by defining $y^3 = V$. The propagation equation for $y(t)$ along $\lambda(t)$ is the Raychaudhuri equation (5.13), which is a second order differential equation now given by

$$\frac{d^2y}{dt^2} + \tfrac{1}{3}(R_{ij}V^iV^j + 2\sigma^2)y = 0, \tag{5.24}$$

where $\sigma^2 = \sigma_{ij}\sigma^{ij}$ corresponds to the trace of the shear tensor σ_{ij}. (For the null case, similar equation holds with 1/3 replaced by 1/2.) Writing $F(t) = \frac{1}{3}(R_{ij}V^iV^j + 2\sigma^2)$ and ignoring the effects of the shear tensor, which will any way enhance the focusing effect we are considering, we choose $F(t) = A/t^2$, where $A > 0$ is a fixed constant. If we try a solution of the form $y = t^\alpha$ then the condition on A is obtained to be $A = \alpha - \alpha^2$. Since $V(t) \to 0$ in the limit of approach to the singularity $t = 0$, we must have $\alpha > 0$. Again, $A > 0$ and the above expression for A implies that $0 < \alpha < 1$. Solving for α gives that in order for α to be real, A satisfies $A \leq 1/4$. The solution for y is then given by

$$y = t^{[1\pm(1-4A)^{1/2}]/2}. \tag{5.25}$$

Thus, depending on the value of A it is seen that $y \sim t^\alpha$ with $\frac{1}{2} \leq \alpha < 1$. The volume $V(t)$ then goes to zero near the singularity at least as fast as $t^{3/2}$.

The criteria on the strength of a singularity are of course subject to further refinement. However, the important physical consequences of the existence of a singularity are related to its strength. The point is if the singularity is gravitationally weak, it may be possible to extend the space-time through the same classically. On the other hand, when there is a strong curvature singularity forming in the above sense, the gravitational tidal forces associated with this singularity are so strong that any object trying to cross it gets destroyed. Thus, as argued by Ori (1992), the extension of space-time becomes meaningless for such a strong singularity which destroys to zero size all the objects terminating at the singularity. From this point of view, the strength of singularity may be considered crucial to the issue of classically extending the space-time and thus avoiding the singularity; because for a strong curvature singularity defined in the above sense, no continuous extension of the space-time may be possible.

6

GRAVITATIONAL COLLAPSE AND COSMIC CENSORSHIP

In this chapter, we investigate the issue of the final fate of a gravitationally collapsing massive star and the associated cosmic censorship problem. When the star is heavier than a few solar masses, it could undergo an endless gravitational collapse without achieving any equilibrium state. This happens when the star has exhausted its internal nuclear fuel which provides the outwards pressure against the inwards pulling gravitational forces. The theory of singularities discussed in Chapter 5 then implies that for a wide range of reasonable initial data, a space-time singularity must develop. The conjecture that such a singularity of gravitational collapse from a regular initial surface must always be hidden behind the event horizon of gravity is called the cosmic censorship hypothesis (Penrose, 1969). Thus, cosmic censorship implies that the final outcome of gravitational collapse of a massive star must necessarily be a black hole which covers the resulting space-time singularity. Then, causal messages from the singularity cannot reach the external observer at infinity and the predictability in space-time is fully preserved. This hypothesis lies at the foundation of currently well-developed and applied theory of black hole physics.

It has not been possible, however, to obtain a proof or to establish the validity of cosmic censorship despite several serious attempts, and this remains one of the most outstanding open issue in gravitation theory today (see for example, the reviews by Eardley, 1987; Israel, 1984, 1986). The point is, if naked singularities could result from realistic gravitational collapse, the possible causal emissions from the same would introduce altogether new possibilities in gravitation physics and the physical relevance of black holes would be seriously undermined.

We discuss in Section 6.1 the essential features of spherically symmetric collapse and the final fate of a massive star which undergoes gravitational collapse when it has exhausted its nuclear fuel. It is the spherically symmetric collapse of a homogeneous dust cloud, as described by the Oppenheimer–Snyder (Oppenheimer and Snyder, 1939) model, which led to the general concept of trapped surfaces and the black hole. The event horizon and other basic features of black holes are discussed in Section 6.2. The expectation is that, even when the collapse is inhomogeneous or non-spherical, the black holes must always form covering the singularity implied by the trapped surface. The evidence in favour of and against such a censorship proposition is reviewed in Section 6.3. The available versions of censorship hypothesis are given there, pointing out formidable difficulties in attempting any possible proof. It is concluded that the first major task is to formulate a provable version of the cosmic censorship conjecture, and that to achieve such a purpose a detailed and careful analysis of available gravitational collapse scenarios is essential where the possible occurrence and physical nature of the naked singularity forming is to be analysed. It is only such an examination of collapse situations which could tell us what features are to be avoided, and which ones to look for, while formulating and proving any reasonable version of censorship hypothesis. On the other hand, this analysis could also lead us to better counter-examples, explaining the nature and occurrence of the naked singularity in gravitational collapse more effectively.

The rest of this chapter, beginning at Section 6.4, is devoted to such an analysis of several important gravitational collapse scenarios. In Section 6.4 we investigate in detail the phenomena of naked singularity formation in the Vaidya–Papapetrou radiation collapse space-times. It is shown that not just isolated radial null geodesics but entire families of future directed non-spacelike geodesics escape from the naked singularity in the past, which forms at the origin of coordinates. The structure of these families is specified in detail and the curvature growth along such trajectories is examined to understand the nature and various features of this naked singularity as fully as possible. It is seen that this is a strong curvature naked singularity (see Section 5.4) in a very powerful sense in that the curvatures diverge very rapidly along *all* the families of non-spacelike geodesics meeting the singularity in the past. It is also shown that the growth of curvatures and the Kretschmann scalar $\alpha = R^{ijkl}R_{ijkl}$ exhibits a directional behaviour in the past near the naked singularity which is similar to that discovered in the cosmological context by Ellis and King (1974).

The radiation models studied in Section 6.4 are a special case of general self-similar space-times where the equation of state is that of an inflowing radiation. These results are generalized in Section 6.5 to study the structure and formation of naked singularities for the general class of self-similar space-times for the collapse of an adiabatic perfect fluid which allows for a non-zero pressure. It is seen that the conclusions on the existence and nature of the naked singularity generalize with the same intensity in that a non-zero measure set of non-spacelike trajectories escape from the naked singularity, which is again a powerfully strong curvature singularity. In Section 6.6 we consider the question of whether the occurrence of strong curvature naked singularities is necessarily confined to self-similar space-times only. The answer is provided in the negative by the analysis of a non-self-similar class of space-times, namely, the family of radiation models with a non-linear mass function. The non-self-similar gravitational collapse of matter with the form of a dust equation of state is investigated in Section 6.7. A general class of non-self-similar Tolman–Bondi space-times is studied there which admit the formation of a strong curvature naked singularity. The Tolman–Bondi dust collapse scenario studied here generalizes earlier conclusions obtained on these models (Eardley and Smarr,1979; Christodoulou, 1984; Newman, 1986a) and shows how the introduction of inhomogeneities cause naked singularities, changing the conclusions obtained from the homogeneous dust collapse picture.

The overall implication from the analysis above is the following. Several reasonable space-times including radiation collapse, dust models and perfect fluid admit naked singularities of a serious nature. This need not be an instantaneous singularity in that families of non-spacelike geodesics could terminate at the same in the past and an observer at infinity can see it for an arbitrary long period of time. Further, it is physically important in the sense of being a strong curvature naked singularity which is gravitationally not weak and in fact implies a very powerful curvature growth. This analysis provides useful input towards formulating a suitable provable version of cosmic censorship. For example, it follows that the censorship hypothesis cannot be true in the form that naked singularities, whenever they form in gravitational collapse, must be gravitationally weak (see for example, Tipler, Clarke and Ellis, 1980; Newman, 1986 a,b; Israel, 1986b). Further implications of this situation towards the possible formulation and proof of cosmic censorship hypothesis and for the crucial issue of final fate of a massive collapsing star are discussed in the next chapter. It appears necessary to deduce possible constraints on naked singularity formation which can be derived in a general way. This issue is taken up in the next chapter, where we indicate certain future directions.

6.1 Collapse of massive stars

In order to understand the final fate of a massive star which has exhausted its nuclear fuel providing the internal pressure against the inwards pull of gravity, we first give an outline of the life of an ordinary star which comes into being by the gravitational contraction of a dense interstellar gas cloud. In the process of such a compression the central temperature of the material rises to ignite a nuclear fuel burning cycle in which the hydrogen, which forms the bulk of the cloud, burns to make helium. The gravitational contraction is halted and the star enters a quasi-static period when it supports itself against gravity by means of the thermal and radiation pressures. Such a phase may continue for billions of years, depending of the original mass of the star. If $M < M_\odot$ (where $M_\odot$ denotes the mass of the sun, $\sim 2 \times 10^{33}$gm), this period is longer than 10^{10} years, but if $M > 10M_\odot$ it has to be less than 2×10^7 years, that is, very massive stars burn out their nuclear fuel much faster. (For a review, see for example, Blandford (1987); Blandford and Thorne (1979), and references therein). The star spends this large portion of its life on the main sequence. The evolution of the star is actually an interplay between the nuclear burning and gravitational collapse and once much of the hydrogen in the core is consumed providing no further pressure against gravity, the collapse must continue further if the star is still sufficiently massive. In the process, the core temperatures rise again to initiate thermonuclear reactions converting helium into carbon, and the core stabilizes again. For a heavy enough star this process repeats itself until a large core of stable nuclei, such as iron and nickel, is built up.

The point is, the final state for such an evolution is either an equilibrium star or a state of continual, endless gravitational collapse. The key factor for such stability analysis is the equation of state for the cool matter of the star in its ground state, that is, when all possible nuclear reactions have taken place and no further energy can be derived from such burning. The support in such a case must come either from the electron degeneracy pressure, when the star becomes a *white dwarf*, or from neutron degeneracy pressure giving a *neutron star* (Fig. 41). Chandrasekhar (1931, 1934) approximated the equation of state in this case by an ideal electron Fermi gas and showed that there is a maximum mass limit for the mass of a spherical, non-rotating star to achieve a white dwarf stable state, which is given by

$$M_c \sim 1.4 \left(\frac{2}{\mu_e}\right)^2 M_\odot, \tag{6.1}$$

(where μ_e is the constant mean molecular weight per electron). Subsequently, considerable work has been done on equations of state for matter at nuclear densities (see for example, Arnett and Bowers, 1977) and it is seen that the maximum mass for non-rotating white dwarfs lies in the range $1.0M_\odot - 1.5M_\odot$ depending on the composition of matter. Similar considerations for neutron stars give this range to be $1.3M_\odot - 2.7M_\odot$ (Arnett and Bowers, 1977).

A typical white dwarf has a radius of $\sim 10^4$km and central density $\sim 10^6 \mathrm{gm\,cm^{-3}}$ whereas for a neutron star these numbers are given by ~ 10 km and $\sim 10^{15}\mathrm{gm\,cm^{-3}}$ respectively. The maximum mass upper limits stated above are raised somewhat when either the rotation or differential rotation of the star is taken into account; however, under very general circumstances a firm upper limit of about $5M_\odot$ is obtained (Hartle, 1978). Many examples of white dwarfs are known to exist in the universe and

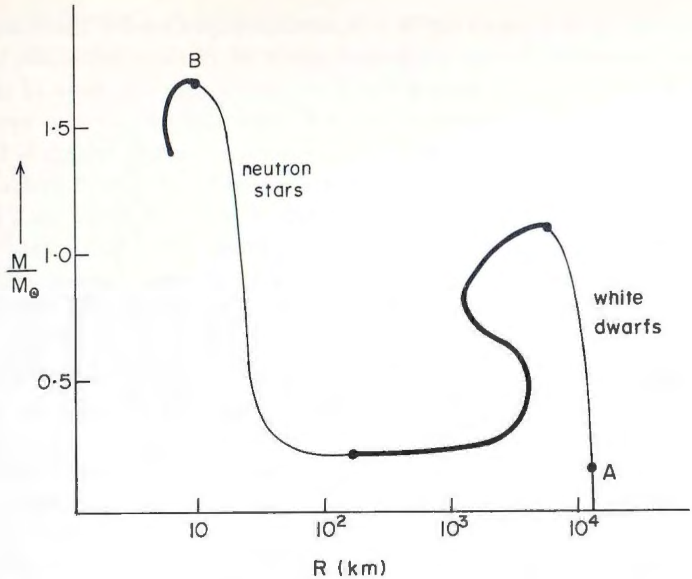

Fig. 41 A schematic diagram for the equilibrium states of massive stars made of cold matter in terms of the masses and the radius of the star. The basic factor determining the equilibrium is the central density ρ of the star. There are a large number of possibilities here as far as the equation of state of the matter is concerned, but for a given equation of state the central density uniquely determines the equilibrium configuration. Each point on the curve corresponds to a single star corresponding to that value of central density chosen to be between 10^5 gm cm^{-3} at A to about 10^{17}gm cm^{-3} at B. The heavier part of the curve denotes unstable region for central densities; the white dwarf configurations are found for lower values of densities compared to the unstable region between A and B and the neutron star region lies towards the higher side. There are no equilibrium states possible beyond point B. In the white dwarf region, the masses and radius depend to an extent on the assumed matter composition of the star, whereas in the neutron star case these depend greatly on the state of the matter as well as the interactions between constituents of matter.

the discovery of pulsars has provided strong support for the existence of neutron stars, which must be rotating with periods of fractions of a second in order to produce the observed pulsar signals.

Hence, if a star has a mass higher than or about $5M_\odot$, it must enter a state of perpetual gravitational collapse once it has exhausted all its nuclear fuel and no equilibrium configurations are possible unless it manages to throw away most of its mass by some process during this evolution. In fact, mass ejection is observed in a *supernova* explosion for the star. When the core collapse is halted or slowed down at nuclear densities, a shock wave is produced which propagates outwards in the envelope of the star. While the inner core remains a neutron star, the outer parts are driven away by the shock releasing enormous mass and energy, which is believed to be a supernova explosion. The theory for such ejection of matter is not well understood, however, and at any rate it seems unlikely that all such massive stars will be able to throw away almost all of their mass in such a process. The reason is, for stars having tens of solar masses, this would amount to throwing away almost ninety percent of the mass of the star. No suitable mechanisms are envisaged today which could achieve such a high degree of efficiency. Thus, if the shock could not blow off all the outer layers, they would fall on the newborn neutron star and the collapse continues again.

We examine in the next section the final fate of such a spherically symmetric massive star in the state of a perpetual gravitational collapse. After describing some general features of spherical collapse, we confine ourselves to the case of a homogeneous dust cloud. The assumption of cosmic censorship then allows us to generalize the concept of a black hole obtained from this specific case when the perturbations from spherical symmetry are taken into account (see Section 6.3). The departures resulting, when (spherically symmetric) inhomogeneities are introduced, will be considered in Section 6.8.

6.2 Spherically symmetric collapse

In order to understand the possible final fate of a massive gravitationally collapsing star, we consider here the spherically symmetric collapse situation. Such a symmetry represents a high degree of idealization of the physical situation but the advantage is that one can solve it analytically to get exact results when matter is taken in the form of a homogeneous dust cloud. It is also possible that many salient features of the situation, including the notion of a black hole discussed in the next section, could be preserved when departures from spherical symmetry are taken into account. In fact, the basic motivation for the idea and theory of black holes comes from this case of homogeneous dust cloud collapse. Independently of the interior solution, the metric exterior to such a spherical body must be the Schwarzschild space-time and no gravitational radiation will be present, which follows from *Birkhoff's theorem* which is that the only vacuum, spherically symmetric gravity field must be static.

In order to consider spherically symmetric space-times, if P is any point at a distance r from the origin O, the system must be invariant under rotations around O. Such rotations will generate a two-sphere around O, the line element on which must be given by

$$ds^2 = r^2(d\theta^2 + \sin^2\theta\, d\phi^2).$$

This is the line element for a two-sphere given by $t = \text{const.}$, $r = \text{const.}$ in a general spherically symmetric space-time. Further, as the metric must be invariant under the reflections $\theta \to \pi - \theta$ and $\phi \to -\phi$, there must not be any cross terms in the metric in $d\theta$ and $d\phi$. As the line element must not change with any change in θ and ϕ, they must occur in the metric only in the form of the two-metric given above. Then, in the (t, r, θ, ϕ) coordinate system, the metric has the form

$$ds^2 = -Adt^2 + 2Bdt\,dr + Cdr^2 + D(d\theta^2 + \sin^2\theta\, d\phi^2).$$

Here the quantities A, B, C, and D are the functions of t and r to be determined. Introducing now a new radial coordinate $r' = D^{1/2}$, and a new time coordinate t' by requiring that $dt' = F[A'dt - B'dr']$, where $F(t, r')$ is a suitable integrating factor, the line element reduces to

$$ds^2 = -e^{\mu}dt^2 + e^{\nu}dr^2 + r^2(d\theta^2 + \sin^2\theta\, d\phi^2),$$

where we have dropped the primes. Here $\mu = \mu(r, t)$ and $\nu = \nu(r, t)$, and the quantities $e^{\mu} = 1/F^2A'$ and $e^{\nu} = C' - B'^2/A'$ appearing in the metric are always positive.

In general, one can define the *spherical symmetry* of a space-time in terms of the Killing vectors; there must be three linearly independent space-like Killing vector fields X^1, X^2 and X^3, in the space-time which satisfy the commutator relations

$$[X^1, X^2] = X^3, \quad [X^2, X^3] = X^1, \quad [X^3, X^1] = X^2,$$

and their orbits must be closed. Using these properties, one could then again derive rigorously the line element above for a spherically symmetric space-time.

Consider now a spherically symmetric massive star, collapsing gravitationally when it has exhausted its internal nuclear fuel. We need to consider the interior solution for such a star. Of course, there is no unique interior solution available which basically depends on the properties of the matter, the equation of state obeyed by the matter, and the physical processes taking place within the steller interior. However, assuming the matter to be the pressureless dust allows one to solve the problem analytically, which provides many important insights. In that case, the energy-momentum tensor is given by $T^{ij} = \rho u^i u^j$ and one needs to solve the Einstein equations for the spherically symmetric form of the metric given above. Solving the Einstein equations determines the metric potentials completely and the interior geometry of the star, which is a collapsing dust ball, is described by the same line element as that of the closed homogeneous and isotropic Friedmann models (Section 3.6) given by

$$ds^2 = -dt^2 + R^2(t)\left[\frac{dr^2}{1-r^2} + r^2 d\Omega^2\right],$$

where $d\Omega^2$ represents the metric on a two-sphere. The geometry outside the star is vacuum and is of necessity the Schwarzschild space-time as implied by the Birkhoff theorem. It is possible to show that the interior geometry of the dust cloud matches correctly at the boundary of the star $r = r_b$ with the exterior Schwarzschild space-time.

When the collapse is complete, the space-time settles to a vacuum Schwarzschild geometry (3.26) for the range of coordinate $0 < r < \infty$ (with a coordinate singularity at $r = 2m$, which can be removed by going to the Kruskal extension). Here, m can be identified with the total mass of the star. To understand the structure of the resulting configuration, we consider the radial null geodesics in this metric defined by $ds^2 = 0$ and $\dot{\theta} = \dot{\phi} = 0$. Taking the positive sign solutions, these are given by

$$\frac{dt}{dr} = \frac{r}{r-2m}.$$

Integrating this gives

$$t = r + 2m \ln \mid r - 2m \mid + \text{const.}$$

In the region $r > 2m$ we have $dr/dt > 0$ and hence r increases with increasing t. Thus the above describes the congruence of outgoing radial null geodesics. The ingoing null trajectories are given by the negative sign solutions, which are given by

$$t = -r - 2m \ln \mid r - 2m \mid + \text{const.}$$

The structure of null geodesics is shown in Fig. 42. In the region below $r = 2m$, the coordinates r and t change their spacelike and timelike nature and hence the light cones tip over. As a result, no observer in the region $r < 2m$ can remain at a constant value of r but must move inwards to

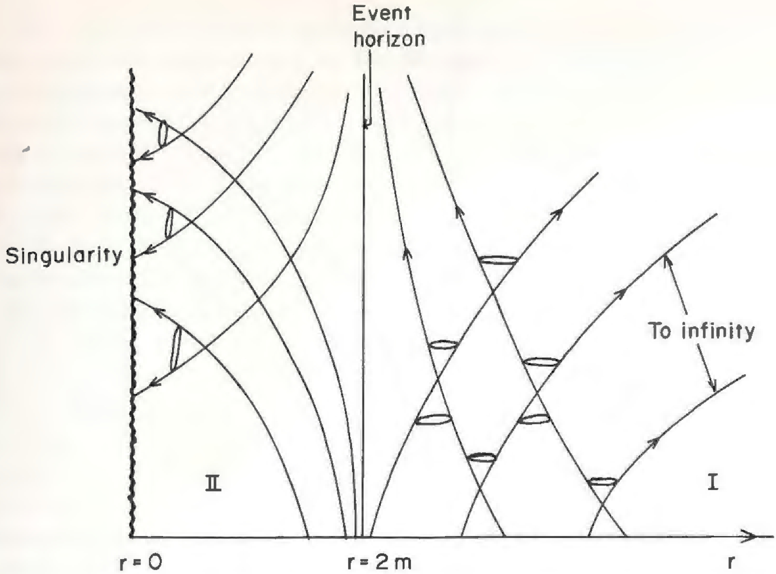

Fig. 42 The structure of null geodesic trajectories in Schwarzschild space-time in Schwarzschild coordinates in (t, r) plane. There is a coordinate singularity at $r = 2m$.

fall within the intrinsic curvature singularity at $r = 0$. Each point in this diagram represents a two-sphere of area $4\pi r^2$.

The above picture gives an idea of the phenomena happening in the region $r < 2m$, namely that any material particle or photon here must fall in the singularity and that it can never escape to larger values of r to communicate with external observers in the space-time. Hence, this may be termed as the black hole region in the space-time. However, the Schwarzschild picture above gives the impression that from the outside space-time $r > 2m$, no photons or particles could fall in this black hole and they will take infinite time to reach the surface $r = 2m$. It turns out that this is just a coordinate defect arising due to the coordinate singularity at $r = 2m$, as can be seen by going to the Eddington–Finkelstein coordinates as below. The idea here is to choose a new time coordinate such that the ingoing null geodesics become straight lines in the space-time. It is clear from the Schwarzschild consideration above that the appropriate change may be given by

$$t \to t + 2m \ln(r - 2m),$$

for $r > 2m$. The solution with such a coordinate change is now regular at $r = 2m$ and in fact, the coordinate range now is, $0 < r < \infty$. This is

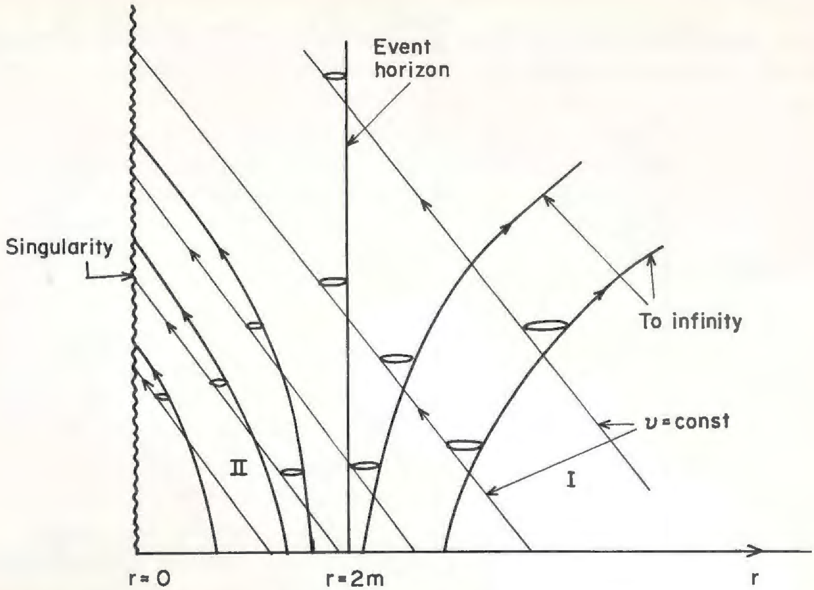

Fig. 43 Schwarzschild solution in advanced Eddington–Finkelstein coordinates.

called an analytic extension of the Schwarzschild solution. A time-reversed solution is obtained if we introduce a different time coordinate

$$t \to t - 2m \ln(r - 2m),$$

in which case, the outgoing null geodesics are straight lines. A simpler way to write the metric now in the new coordinate system is to introduce the advanced null coordinate v defined by

$$v = t + r + 2m \ln(r - 2m).$$

The metric then has the form

$$ds^2 = -\left(1 - \frac{2m}{r}\right) dv^2 + 2dv\,dr + r^2(d\theta^2 + \sin^2\theta\, d\phi^2).$$

The radially infalling null geodesics are now given by $v = \text{const.}$, as seen in the space-time diagram given in Fig. 43. It can be seen that at $r = 2m$, the radially outgoing photons stay at a constant value of r and below this surface, they also must fall to the singularity at $r = 0$. Further, radially infalling material particles also must fall to the space-time singularity within a finite amount of proper time as measured along their trajectory.

The basic features of a collapsing spherically symmetric dust cloud configuration are summarized by the Penrose diagram given in Fig. 15b. The collapse is initiated when the star surface is outside its Schwarzschild radius $r = 2m$ and a light ray emitted from the surface of the star can escape to infinity. However, once the star has collapsed below $r = 2m$, a *black hole*, that is, a region of no escape develops in the space-time which is bounded by the event horizon at $r = 2m$. Any point in this empty region in the Kruskal diagram represents a *trapped surface* (which is a two-dimensional sphere in space-time) in that both the outgoing and ingoing families of null geodesics emitted from this point converge and hence no light ray comes out of this region bounded by $r = 2m$. Then, the collapse to an infinite density and curvature singularity at $r = 0$ becomes inevitable in a finite proper time as measured by an observer on the surface of the star. Then, the black hole region in the resulting vacuum Schwarzschild geometry is given by $0 < r < 2m$ and the outer boundary of this region, namely, $r = 2m$ is called the *event horizon*. On the event horizon, only the radial outwards photons stay where they are but all the rest are dragged inwards. No information from this black hole region can propagate outside $r = 2m$ to an outside observer. An alternative picture of this gravitational collapse phenomena for the homogeneous dust ball and the resulting black hole is given in Fig. 1.

The general formalism for a spherically symmetric gravitational collapse, including pressure, was given by Misner and Sharp (1964). The spherically symmetric space-time can be written in co-moving coordinates as

$$ds^2 = -e^{2\phi}dt^2 + e^{\lambda}dr^2 + R^2(t,r)d\Omega^2, \tag{6.2}$$

where $d\Omega^2 = d\theta^2 + \sin^2\theta\, d\phi^2$ is the usual metric on the two sphere and ϕ and λ are functions of t and r. The stress-energy tensor is that of a perfect fluid given by

$$T_{ij} = (\rho + p)u_i u_j + p g_{ij}. \tag{6.3}$$

The spatial velocities u^i are vanishing here and the spatial coordinates of a given particle remain constant throughout the collapse. A function $m(r,t)$ is introduced by the definition

$$e^{\lambda} = \left(1 + \dot{R}^2 - \frac{2m}{r}\right)^{-1} R'^2, \tag{6.4}$$

where a dash denotes derivative with respect to r and for any function f,

$$\dot{f} = e^{-\phi}\left(\frac{\partial f}{\partial t}\right). \tag{6.5}$$

The coordinate t gives proper time along the particle world lines. Integrating the conservation equation $T^{ij}{}_{;j} = 0$ and solving the Einstein equations, the Misner–Sharp equations for the spherically symmetric collapse are written as follows:

$$\dot{m} = -4\pi R^2 p \dot{R},$$
$$\ddot{R} = \left(\frac{1 + \dot{R}^2 - 2m/r}{\rho + p}\right)\left(\frac{\partial p}{\partial R}\right) - \frac{m + 4\pi R^3 p}{R^2}, \tag{6.6}$$
$$\frac{\partial m}{\partial R} = 4\pi R^2 \rho.$$

The above, when combined with an equation of state relating ρ and p, determine the dynamical evolution of the collapse. However, when the pressure $p \neq 0$, the situation is quite complex and it is necessary to use numerical computation (see for example, May and White, 1966) to get any idea of the evolution of the collapsing system.

6.3 Black holes

In the previous section, we examined the final fate of a spherically symmetric massive star which collapses gravitationally. General relativity predicts that the star contracts until all the matter in the star collapses to a space-time singularity. This happens when the star is sufficiently massive so that no equilibrium state is possible and the gravitational pull overcomes all the internal pressures and stresses which might possibly stop the collapse. This gives rise to a black hole in the space-time which covers the space-time singularity resulting from the collapse.

In this section we consider certain important general properties of such black holes. For a detailed treatment on black hole physics and results such as black hole uniqueness theorems and so on, we refer to Hawking and Ellis (1973), Hawking and Israel (1979), and references therein. As stated earlier, the fundamental motivation for the concept of a black hole comes from an examination of the spherically symmetric homogeneous collapse which has two outstanding features. Firstly, for a star undergoing a complete gravitational collapse, a region of trapped surfaces form below $r = 2m$, from which no light rays escape to an observer at infinity. Thus, a black hole forms in the space-time. Secondly, the ultimate fate of the star undergoing the collapse is an infinite curvature singularity at $r = 0$, which is completely hidden within the trapped surface region and the black hole. Hence no emissions or light rays from the singularity could go out to any observer at infinity and the singularity is causally disconnected from the outside space-time.

The question now is whether we can generalize these conclusions for a non-spherically symmetric collapse, and whether they are valid at least for small perturbations from exact spherical symmetry. It has been shown (Hawking and Ellis, 1973), using the stability of Cauchy development in general relativity, that the formation of trapped surfaces, and hence of a black hole, is indeed a stable property when departures from spherical symmetry are taken into account. The argument essentially is the following: Considering a spherically symmetric collapse evolution from given initial data on a partial Cauchy surface S, we find the formation of trapped surfaces T in the form of all the spheres with $r < 2m$ in the exterior Schwarzschild geometry. The stability of Cauchy development then implies that for all initial data sufficiently near to the original data in the compact region $J^+(S) \cap J^-(T)$, the trapped surfaces must still occur. Then, the curvature singularity of spherical collapse also turns out to be a stable feature as implied by the singularity theorems, which show that the closed trapped surfaces always imply the existence of a space-time singularity under reasonable general conditions.

There is no proof available, however, that the singularity will continue to be hidden within the black hole and remain causally disconnected from outside observers, even when the collapse is not exactly spherical or when departures from the exact homogeneous dust cloud situation are considered. If the singularity became visible to external observers, the predictability in the space-time will be seriously undermined because arbitrary new information could come from the singularity where the densities and curvatures could be arbitrarily large.

Thus, in order to generalize the notion of black holes to gravitational collapse situations other than exactly spherically symmetric homogeneous dust case, it becomes necessary to rule out such naked singularities by means of an explicit assumption. This is stated as the *cosmic censorship hypothesis*, which says that if S is a partial Cauchy surface, there are no naked singularities to the future of S, that is, which can be seen from the future null infinity $\mathcal{I}^+$. This is true for the spherical homogeneous dust cloud collapse where the breakdown of physical theory at the space-time singularity does not disturb the prediction in future for the outside asymptotically flat region.

Following Hawking and Ellis (1973), we make this precise by assuming the framework of space-times which admit weakly asymptotically simple and empty conformal completions (Section 4.7). These space-times are highly appropriate to model collapse scenarios of isolated objects and include the Schwarzschild, Reissner–Nordström and Kerr solutions. Suppose now (M, g) is a space-time admitting a weakly asymptotically simple and empty conformal completion $(\overline{M}, \bar{g})$. Then we say that (M, g) is *future* *asymptotically predictable* from a partial Cauchy surface S if

$$\mathcal{I}^+ \subset \bar{D}^+(S, \overline{M}),$$

that is, the future null infinity $\mathcal{I}^+$ is contained in the closure of $D^+(S)$ in the conformal manifold.

Future asymptotic predictability ensures the cosmic censorship condition in the form that there are no singularities in the future of S which are 'naked', that is, visible from the future null infinity $\mathcal{I}^+$. In the spherical homogeneous dust collapse, the resulting space-time is future asymptotically predictable and the censorship holds. What will be the situation for a non-spherical and other collapse situations? Answer to this question is not known either in the form of a proof for the future asymptotic predictability for general space-times or of any other suitable version of cosmic censorship hypothesis. The evidence in support of this conjecture will be reviewed in the next section, which forms a basic hypothesis for the black hole physics.

A *black hole region* in the space-time, denoted by $\mathcal{B}$, for a future asymptotically predictable space-time M is defined as

$$\mathcal{B} = M - J^-(\mathcal{I}^+).$$

Thus, this is a space-time region from which no null or timelike curves can reach an observer at infinity (Fig. 44). The boundary of $\mathcal{B}$ in M, given by

$$\mathcal{H} = \dot{J}^-(\mathcal{I}^+) \cap M,$$

is called the *event horizon*. As pointed out in Section 4.2, this horizon must be an achronal surface generated by null geodesics which could have past end points in M but have no future end points. For the Minkowski space-time, $J^-(\mathcal{I}^+) = M$ and there is no black hole; however, for the Schwarzschild case, $J^-(\mathcal{I}^+)$ is the region for space-time exterior to $r = 2m$ and the event horizon is given by the null hypersurface $r = 2m$ which is the boundary of the black hole region $0 < r < 2m$.

As noted above, no space-time singularities are visible at null infinity in an asymptotically predictable space-time. In fact, this is true for trapped surfaces also in this case, provided either weak or strong energy condition is satisfied. This is seen from the following result (for the proof we refer to Hawking and Ellis (1973)), which states that all trapped surfaces must be fully contained within the black hole region, not visible from $\mathcal{I}^+$.

Theorem 6.1. Let (M, g_{ij}) be future asymptotically predictable from a partial Cauchy surface S and $R_{ij}K^iK^j \geq 0$ for all null vectors K^i. If T is a closed trapped surface in $D^+(S)$, then $T \cap J^-(\mathcal{I}^+) = \emptyset$

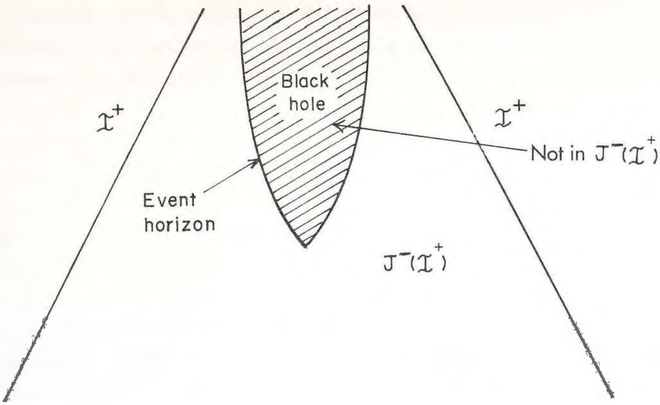

Fig. 44 The black hole is a region of space-time from which no causal communications, that is, non-spacelike curves can reach the future null infinity $\mathcal{I}^+$. The event horizon is the boundary of the region $J^-(\mathcal{I}^+)$ in M.

Let (M, g_{ij}) be an asymptotically flat space-time with the associated unphysical conformal space-time $(\overline{M}, \bar{g}_{ij})$. Suppose $p \in \mathcal{I}^+$ and $q \in M \cap J^-(p)$. Let γ be the future directed null geodesics generator of $\mathcal{I}^+$ through p and let $r \in \gamma$ be any point. Then $q \in I^-(r)$ in M. Thus $J^-(\mathcal{I}^+) = I^-(\mathcal{I}^+)$ in M and hence $J^-(\mathcal{I}^+)$ is open in M. Thus the black hole region $\mathcal{B} = M - J^-(\mathcal{I}^+)$ is closed in M. This implies that the event horizon is contained in $\mathcal{B}$. As such the black hole region $\mathcal{B}$ need not be connected in M and while considering isolated black holes forming out of gravitational collapse in M, one works with a connected component of $\mathcal{B}$.

Any event p on the event horizon $\mathcal{H}$, however, lies on the boundary of the black hole region and hence any smallest perturbation could make p enter $J^-(\mathcal{I}^+)$, that is, causally connected to the infinity. Then, the space-time is no longer asymptotically predictable. This situation is avoided by further demanding that for the partial Cauchy surface S,

$$J^+(S) \cap \overline{J^-(\mathcal{I}^+)} \subset D^+(S).$$

This effectively means that a neighbourhood of the event horizon could also be predicted from S, and is equivalent to the condition that the space-time exterior to the black hole region is globally hyperbolic.

In the case of collapsing dust, the event horizon is the null hypersurface generated by those null geodesics which just reach the surface of the star when it crosses the radius $r = 2m$ with m being the Schwarzschild mass of the star. The area of the horizon increases monotonically until the horizon reaches surface of the star. Outside, this area is a constant given by $A = 16\pi m^2$. For the Kerr space-time written in the Boyer–Lindquist form (3.45) the horizon is defined by $r = r_+$ given by eqn (3.49). The area is obtained by setting $t =$ const., $r = r_+$ which gives the metric on the surface. Then we have

$$A = \int \sqrt{g}\, d\theta d\phi = 8\pi m[m + (m^2 - a^2)^{\frac{1}{2}}]. \tag{6.7}$$

Thus, the area of the horizon is a non-decreasing function.

In fact, for strongly asymptotically predictable space-times in general the area of a black hole horizon must either remain constant or must increase provided $R_{ij}K^iK^j \geq 0$ for all null vectors K^i (Hawking, 1971). The basic argument leading to the proof of this result is the following, which describes the evolution of the event horizon. First we note that the horizon $\mathcal{H}$ is generated by future inextendible null geodesics generators, because $\mathcal{H}$ is the boundary of the past of $\mathcal{I}^+$ and so in order to have future endpoint, a generator must interest $\mathcal{I}^+$ which is not possible. Next, for all the null geodesic generators of $\mathcal{H}$, the expansion θ must be everywhere non-negative, $\theta \geq 0$. This is because, if at some $p \in \mathcal{H}$, if $\theta < 0$, one could deform $\mathcal{H}$ in the neighbourhood of p so that $\theta < 0$ still in this neighbourhood, and it enters the past of $J^-(\mathcal{I}^+)$. Now, choose a spacelike two-surface T in $\dot{J}^-(\mathcal{I}^+)$ in this neighbourhood, then T intersects $J^-(\mathcal{I}^+)$. Then, generators of $\dot{J}^+(T)$, visible from $\mathcal{I}^+$ have past endpoints at T and are orthogonal to T. However, $\theta < 0$ implies that they must have a point conjugate to T within a finite affine distance and so cannot remain in the null boundary all the way to infinity, which is a contradiction. Thus, the area of a two-dimensional cross section of generators cannot decrease as $\mathcal{H}$ evolves to the future. From this it is then possible to deduce the result that the area of the event horizon must be non-decreasing in future.

In particular, if two or more black holes merge to form a single black hole, the area of its boundary must be greater than or equal to the sum of the original black hole areas. An interesting implication of this result was pointed out by Hawking (1971) to get an upper limit on the energy that can be emitted in gravitational radiation when two black holes coalesce. Consider two black holes with masses m_1 and m_2 and angular momenta a_1 and a_2 which collide to give a third black hole with these parameters m_3, a_3. The area of a single hole is given by eqn (6.7) and hence the area theorem above implies

$$m_3[m_3 + (m_3^2 - a_3^2)^{\frac{1}{2}}] \geq m_1[m_1 + (m_1^2 - a_1^2)^{\frac{1}{2}}] + m_2[m_2 + (m_2^2 - a_2^2)^{\frac{1}{2}}]. \tag{6.8}$$

The energy radiated is $m_1 + m_2 - m_3$, or the fraction of total energy radiated is

$$f = 1 - \frac{m_3}{m_1 + m_2}. \tag{6.9}$$

Then, using the inequality (6.8) it can be seen that $f < 1/2$, that is, at most half the initial energy could be released in black hole collisions.

Most astrophysical bodies have been observed to possess some rotation and just like a rotating star, a rotating Kerr black hole also must have rotational energy. Various methods have been suggested to extract this rotational energy, which include the method suggested by Penrose (1969), where the existence of the ergosphere is used for energy extraction, and the tidal friction method of Hawking and Hartle (1972). However, through whatever method the rotational energy is extracted, the area theorem places certain general limitations here. In any process of energy extraction the black hole area must be non-decreasing and, following Christodoulou (1970), one could introduce the concept of irreducible mass m_0 for a black hole by the relation

$$A = 16\pi m_0^2. \tag{6.10}$$

Then, using the expression (6.10) for the area one can write

$$m^2 = m_0^2 + \frac{J^2}{4m_0^2}, \tag{6.11}$$

where J is the angular momentum of the black hole. The second term in the above equation can be interpreted as the rotational contribution to the black hole mass. One could extract a maximum of rotational energy given by $m - m_0$. An important physical process that must be taken into account while considering the energy extraction is the accretion of the surrounding matter by the black hole (Blandford and Thorne, 1979) and the inherent possibility that this might increase the rotational energy of the black hole.

According to the area theorem above, for all physically allowed processes, the total area of black holes cannot decrease, that is, $\delta A \geq 0$. This is very similar to the second law of thermodynamics, which says that for all physical processes, the total entropy of all the matter in universe is non-decreasing, that is, $\delta S \geq 0$. Consider, for example, the Schwarzschild black hole. The area for its event horizon is given by $A = 16\pi m^2$. The only way to reduce this area is to extract mass from the black hole, which is impossible because no particles or photons can cross the event horizon to come out. On the other hand, one could increase the area by throwing particles in which would increase the mass (see for example, Bekenstein (1973) for details on the relationship between entropy and the area of event horizon). Similar analogies have been developed between other laws of thermodynamics and the laws of black hole physics. In fact, it has been shown by Bardeen, Carter and Hawking (1973) that for any stationary axisymmetric black hole in an asymptotically flat space-time, it is possible to define a quantity called surface gravity which is constant on the horizon. Thus, just as the temperature is constant throughout a body in thermal equilibrium (zeroth law of thermodynamics), so is the surface gravity over the horizon of a stationary black hole. Again, for a Schwarzschild black hole, the surface gravity turns out to be $1/4m$ and it is impossible to reduce it to zero by any physical process just as it is impossible to achieve the zero temperature for a system by any physical process (the third law of thermodynamics). As such, the thermodynamic temperature of a black hole in classical relativity will be absolute zero since it is a perfect absorber which does not emit at all. So it would appear that the surface gravity may not represent physical temperature. However, it was shown by Hawking (1975) that when quantum particle creation effects are taken into account, a black hole actually radiates with a black body spectrum at a temperature proportional to the surface gravity. In this sense, the surface gravity represents the thermodynamic temperature of a black hole. Quantum effects in a black hole geometry will be discussed further in Chapter 9.

The results on black holes discussed so far have assumed the space-time to be asymptotically flat. In reality, however, the universe is probably not asymptotically flat in view of the observed distribution of matter at the largest possible scales. Hence, it would be desirable to have the laws of black holes given in a more general space-time framework. Black holes in a general globally hyperbolic space-time have been defined by Tipler (1977c) and Joshi and Narlikar (1982), who examined the laws of black hole physics in such a framework. It is shown there that for a closed globally hyperbolic space-time with a compact Cauchy surface and a strong curvature crushing singularity in future, the sum of the areas of black holes must decrease in future. It is thus seen that in the situation of the space-time not being asymptotically flat, the behaviour of black holes and the laws governing them are open to further examination.

It is worth noting here that in fact, apart from general relativity, there is a motivation to consider such black holes in the Newtonian theory as well in the sense that the gravitational field of a star could be so strong that it might stop even light (considered as particles) from escaping. For example, consider a spherically symmetric mass distribution of uniform density with radius R and total mass M. For a particle with a velocity v away from the center and the distance r from the center, its total energy is conserved, which is the sum of its kinetic and potential energies. Suppose now the velocity v_0 of the particle is such that it is able to escape to infinity where it has a vanishing velocity. Since the total energy will be zero at $r = \infty$ with $v = 0$, which is conserved, we can write

$$v_0^2 = \frac{2GM}{R}.$$

When the radial velocity of the particle is less than v_0, it must fall back to the body, otherwise it escapes to infinity. Thus, if the mass distribution and radius of the body were such that $c^2 = 2GM/R$, where c is the velocity of light, then for any larger mass or smaller radius of the body, even light will not escape from the body. This was realized by Laplace in 1798, who pointed out that for a star with the density same as the sun but radius 250 times larger, no light could escape from its surface.

What is the observational evidence for the existence of black holes? It is clear that a black hole could not be observed directly, but we must look for the gravitational effects exhibited by such an object. Though there is no conclusive evidence available for the existence of black holes at the moment, presently the best candidates seem to be the binary stars in which one of the partners is visible and the other is supposed to be a black hole. Such a black hole would suck matter from its visible component, in the process forming an accretion disc around the black hole. Before the infalling matter spirals down the black hole, the inner hot regions are believed to produce intense bursts of X-rays formed by synchrotron radiation. Thus, the discovery of the X-ray source Cygnus XI in 1971, which shows rapid variations, indicates the likely evidence for black holes. Further to this, several other X-ray binaries have also been proposed as possible candidates for black holes.

6.4 The formulation of cosmic censorship

It is clear from considerations in Section 6.2 that the assumption of cosmic censorship, in the form of either asymptotic predictability or strong asymptotic predictability of the space-time, is crucial to basic results in black hole physics. In fact, when one considers the gravitational collapse in a generic situation, the very existence of black holes requires this hypothesis (Penrose, 1969).

If one is to establish the cosmic censorship by means of a rigorous proof, that of course requires a much more precise formulation of this hypothesis. The statement that the result of a complete gravitational collapse must always be a black hole and not a naked singularity, or all singularities of collapse must be hidden in black holes, causally disconnected from observers at infinity, is not rigorous enough. This is because, under completely general circumstances, the censorship or asymptotic predictability is false as one could always choose a space-time manifold with a naked singularity which would be a solution to Einstein's equations if we define

$$T_{ij} \equiv \frac{1}{8\pi} G_{ij}.$$

Thus, at the minimum, certain conditions on the stress-energy tensor are required, for example, an energy condition. However, it turns out that to

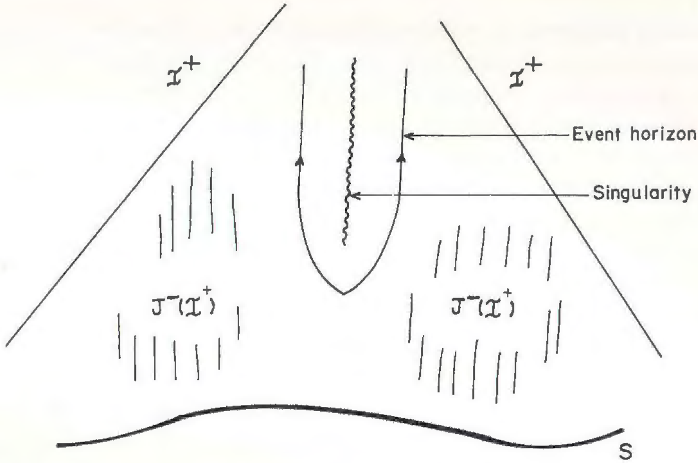

Fig. 45 Asymptotic predictability in the space-time. The points arbitrarily near $\mathcal{I}^+$ must lie in the future domain of dependence of the partial Cauchy surface S.

obtain an exact characterization of the restrictions one should require on matter fields in order to prove a suitable version of the cosmic censorship hypothesis is an extremely difficult task and no such specific conditions are available presently.

The requirements in the black hole physics and general predictability requirements in gravitation theory has led to several different formulations of cosmic censorship hypothesis. The version known as the *weak cosmic censorship* refers to the asymptotically flat space-times and has reference to the null infinity. Weak censorship, or asymptotic predictability, effectively postulates that the singularities of gravitational collapse cannot influence events near future null infinity $\mathcal{I}^+$. If S is the partial Cauchy surface on which the regular initial data for collapse is defined, this is the requirement that $\mathcal{I}^+$ is contained in the closure of $D^+(S)$ (Fig. 45). Thus, the data on S predicts the entire future for far away observers. The other version, called the *strong cosmic censorship*, is a general predictability requirement on any space-time, stating that all physically reasonable space-times must be globally hyperbolic (see for example, Penrose (1979), Geroch and Horowitz (1979) and Hawking (1979)).

In effect, the weak cosmic censorship or the strong asymptotic predictability requirement states that the region of space-time outside a black hole must be globally hyperbolic (Tipler, Clarke and Ellis, 1980). A precise formulation of this version of censorship will consist in specifying exact conditions under which the space-time would be strongly asymptotically predictable. In its weak form the censorship conjecture does not allow causal influences from singularity to asymptotic regions in space-time to an observer at infinity, that is, the singularity cannot be globally naked. However, it could be locally naked in the sense that an observer within the event horizon and in the interior of the black hole could possibly receive particles or photons from the singularity. Thus, one could formulate the weak censorship hypothesis as below (see for example, Wald, 1984):

Let S be a complete spacelike hypersurface on which a generic non-singular initial data (h_{ij}, χ_{ij}) is given in the form of the induced metric and its time derivatives on S. The matter sources are such as to satisfy the dominant energy condition and physically reasonable equations of state. Then, the evolution of gravitational collapse from S must be such that the resulting space-time (M, g_{ij}) is strongly asymptotically predictable.

Clearly, one would like to sharpen the above formulation in many ways. For example, the metric on S should approach that of Euclidian three-space at infinity and matter fields should satisfy suitable fall off conditions at spatial infinity; also one might want the null generators of $\mathcal{I}^+$ to be complete, and one has to specify what exactly is meant by 'physically reasonable' matter fields. In fact, as far as the cosmic censorship hypothesis is concerned, it is a major problem in itself to find a satisfactory and mathematically rigorous formulation of what is physically desired to be achieved (see for example, Penrose, 1982). Developing a suitable formulation would probably be a major advance towards the solution of the main problem.

Since we are interested in the gravitational collapse scenario in this chapter, we require that the space-time contains a regular initial spacelike hypersurface on which the matter fields, as represented by the stress-energy tensor T_{ij}, have a compact support and all physical quantities are well-behaved on this surface. Also, we generally require that the matter satisfies a suitable energy condition, and that the Einstein equations are satisfied. Then we say that the space-time contains a naked singularity if there is a future directed non-spacelike curve which reaches a far away observer or infinity in future, and in the past it terminates at the singularity.

What is the evidence available in favour of the weak cosmic censorship as stated above? Firstly, it should be noted that presently no general proof is available for any suitably formulated version of the weak censorship. The main difficulty appears to be that the event horizon is a feature depending on the whole future behaviour of the solution over an infinite time period, whereas the present theory of quasi-linear hyperbolic equations guarantee the existence and regularity of the solutions over a finite time internal only (Israel, 1984). In this connection the results of Christodoulou (1986) on generic spherically symmetric collapse of massless scalar fields are relevant, where it is shown using global existence theorems on partial differential equations that global singularity free solutions exist for weak enough initial data. In any case, even if it is true, the proof for a suitable version of the weak censorship conjecture would seem to require much more knowledge on general global properties of Einstein's equations and solutions than it is known presently.

While there is no direct proof available at the moment, there is some indirect evidence available for the weak censorship in the form of perturbation and computer calculations which seems to be supporting this conjecture. For example, the perturbation calculations by Doroshkevich, Zel'dovich and Novikov (1966) and Price (1972) indicate that small departures from spherical symmetry may not create naked singularities. Also, numerical studies of linear perturbations of Kerr black hole indicate that it is stable (Press and Teukolsky, 1973) under such perturbations. Further support for censorship comes from the fact that a number of attempts to obtain contradictions to certain inequalities for the area of horizon, based on cosmic censorship, have failed (Gibbons, 1972; Jang and Wald, 1977; Gibbons *et al.* 1983). This supports the possibility that collapse always produces a black hole rather than a naked singularity. The basic chain of argument here is the following. Suppose the collapse has reached a stage when a closed trapped surface T develops on a regular spacelike hypersurfaces S. Weak censorship then implies that an event horizon $\mathcal{H}$ must develop and T is contained within $\mathcal{H}$. One could arrange so that the area A of T is smaller than the area of the horizon $A(\mathcal{H})$, $A < A(\mathcal{H})$. Next, as implied by the area theorem, $A(\mathcal{H}) < A(\mathcal{H})\mid_{t=\infty}$. From eqn (6.7),

$$\begin{aligned} A(\mathcal{H})\mid_{t=\infty} &= 8\pi m_\infty[m_\infty + (m_\infty^2 - (J_\infty/m_\infty)^2)^{\frac{1}{2}}] \\ &< 16\pi m^2, \end{aligned}$$

where m is the gravitational mass measured at the spatial infinity on S and the field is supposed to settle to a Kerr geometry according to the uniqueness theorems of Israel, Carter, Hawking and Robinson (see for example, Carter, 1979) after radiating a positive amount of energy $m - m_\infty$ at infinity with $m_\infty < m$. The point is, the consideration here involves only the initial data on S and so Penrose suggested that to obtain counter-examples to censorship one should look for initial data that violate this inequality.

The evidence for weak censorship stated above is mostly indirect and of indicative character and it has not enabled us so far to arrive at a provable rigorous formulation of the weak censorship conjecture. It would seem that we still do not have sufficient data and information available on the various possibilities present for gravitationally collapsing configurations so as to decide one way or the other on the issue of the censorship hypothesis. What appears really necessary is a detailed examination of collapse scenarios other than the exact homogeneous dust collapse case and to examine the possibilities arising in order to have insights into the issue of the final fate of gravitational collapse. Thus, for example, *shell-crossing* naked singularities have been shown to occur in spherical collapse of perfect fluids (Yodzis, Seifert and Müler zum Hagen 1973, 1974), where shells of matter implode in such a way that fast moving outer shells overtake the inner shells, producing a globally naked singularity outside the horizon. These are the singularities where shells of matter pile up to give two-dimensional caustics and the density and some curvature components blow up. The general point of view is, however, that such singularities need not be treated as serious counterexamples to censorship hypothesis because these are merely consequent to intersection of matter flow lines. This gives only a distributional singularity which is gravitationally weak in the sense that curvatures and tidal forces remain finite near the same (Hawking 1979; Tipler, Clarke and Ellis, 1980).

On the other hand, there are *shell-focusing* naked singularities occurring at the center of spherically symmetric collapsing configurations of dust or perfect fluid or radiation shells, as we shall show in the following sections. These are more difficult to ignore. One can rule them out only by saying that the dust or perfect fluid are not really 'fundamental' forms of matter. However, if the cosmic censorship is to be established as a rigorous theorem, this objection has to be made precise in terms of a clear and simple restriction on the stress-energy tensor, because these are forms of matter which otherwise satisfy reasonability conditions such as the dominant energy condition (provided there are no large negative pressures) or a well posed initial value formulation for the coupled Einstein-matter field equations. Further, these forms of matter are widely used in discussing various astrophysical processes. A detailed review of the situation to 1980 is given by Tipler, Clarke and Ellis (1980) and for later developments we refer to Israel (1984), Eardley (1987), Newman and Joshi (1988), and Joshi and Dwivedi (1992a,c).

We now discuss briefly the hypothesis of *strong cosmic censorship* (Penrose 1979). Unlike the weak conjecture, the strong version demands that the singularities should not be visible, even to the observers within the black hole, that is, they cannot be even locally naked but are always spacelike and the space-time must be globally hyperbolic (Fig. 46). Thus, unless actually encountered, the observer never sees the singularity. The argument given in favour of such a strong principle is that if cosmic censorship is really a basic principle of nature, there should not be any special role given to the observer at infinity because physical laws operate at a local level. Again, this principle is to be carefully formulated because by suitable cuts and identifications in Minkowski space-time one can generate an inextendible non-globally hyperbolic space-time. For certain proposed formulations of strong censorship and difficulties encountered, we refer to Penrose (1979). Again, no general proof is available for strong censorship but there is some supportive evidence available in the form that Cauchy horizons forming as a result of non-global hyperbolicity of space-time turn out to be unstable in certain cases. Thus, for example, in the Reissner–Nordström case the Cauchy horizon exhibits a so called 'infinite blue-shift' instability (Chandrasekhar and Hartle, 1982). However, again the evidence is not uniform here. For example, Morris, Thorne and Yurtsever (1988) have shown that for the wormhole space-time they have constructed, the Cauchy horizon is immune to the Taub–NUT type of instability and they conjecture that it is fully stable providing a counter-example to strong censorship. An additional condition that is often required in both the weak and strong formulations of censorship is that *stable* space-times do not admit naked singularities. This requires a suitable criteria for stability of space-times which is again a major difficult problem in general relativity.

Certain other formulations of the censorship principle have also been proposed, based on different motivations. Thus, a class of censorship conjectures suggest that all naked singularities must be in some sense gravitationally weak (Tipler, Clarke and Ellis, 1980; Newman, 1986b; Israel, 1986b). We shall discuss this version and related aspects in more detail later in this chapter. Then, in order to avoid the difficulties associated with the question as to what forms of matter and equations of state should be considered reasonable, it is also suggested to examine first a purely vacuum version of censorship. That will tell whether pure gravity allows naked singularities. In fact, Geroch and Horowitz (1979) have detailed several possible approaches to censorship formulation and pointed out difficulties also in each case.

It is now possible to summarize the overall situation as follows. Clearly, the cosmic censorship hypothesis is a crucial assumption underlying all of the black hole physics and gravitational collapse theory, and several important related areas in gravitation theory. Whereas no proof for this conjecture is available, the first major task to be accomplished here is in fact to formulate rigorously a satisfactory version of this hypothesis. The proof of cosmic censorship would confirm the already widely accepted and applied theory of black holes, while its overturn would throw the black hole dynamics into serious doubt. Thus, cosmic censorship turns out to be one of the most important open problem of considerable significance for general relativity and gravitation theory today. Even if it is true, a proof for this conjecture does not seem possible unless some major theoretical advances by way of mathematical techniques and understanding the global structure of Einstein equations are made. In fact, the direction of theoretical advances needed is not quite clear. This situation leads us to conclude that

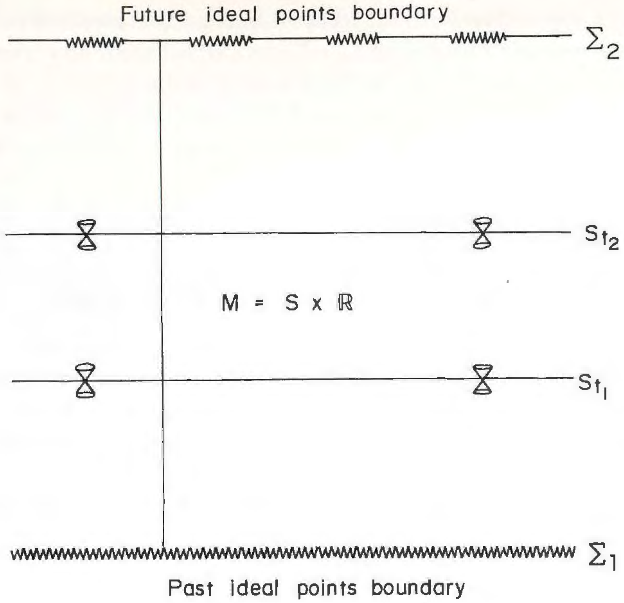

Fig. 46 A space-time obeying the strong cosmic censorship hypothesis, which must be topologically $M = S \times \mathbb{R}$ and globally hyperbolic. The space-time may contain many singularities contained within event horizons due to local black hole formations, which have all been thrown at Σ_2 which is the future ideal points boundary (see Chapter 4). The surface Σ_1 denotes the past ideal points boundary and we note that the topologies of Σ_1 and Σ_2 would be quite different in general, as Σ_2 contains many singularities formed due to local collapse.

the first and foremost task at the moment is to carry out a detailed and careful examination of gravitational collapse scenarios which possibly give rise to a naked singularity formation. It would be fair to say that so far much attention has been paid only to the homogeneous dust collapse and to the implications it suggests, but alternative collapse scenarios have not been given the required attention. Until this is done, trying out different formulations for censorship may not probably help because without really knowing what is involved, we might be in for a complete surprise as far as the final fate of collapse is concerned. It is only such an investigation of more general collapse situations which could indicate what theoretical advances to expect for a proof and what features to avoid while formulating the cosmic censorship.

The rest of this chapter is devoted to such an investigation of several general classes of gravitational collapse scenarios which include models such as the Vaidya–Papapetrou radiation collapse, Tolman–Bondi inhomogeneous dust space-times, general self-similar collapse of a perfect fluid, and also certain non-self-similar situations. The implications of this analysis towards the formulation and proof of cosmic censorship are discussed as and when relevant and also in the next chapter, where possible future directions are indicated.

6.5 Censorship violation in radiation collapse

We begin our study of gravitational collapse scenarios, other than the special case of completely homogeneous dust cloud collapse, by investigating the collapse of inflowing radiation. This is the situation in which a thick shell of radiation collapses at the center of symmetry in an otherwise empty universe which is asymptotically flat far away. The situation discussed in this section could be relevant to the gravitational collapse of a massive star because in the very final stages of the collapse the collapsing matter would be largely radiation dominated. Our main purpose here is to examine the final fate of such a collapse with special reference to the occurrence of naked singularities and the cosmic censorship hypothesis.

When should one regard a naked singularity forming in a gravitational collapse as a serious situation which must guide the formulation and proof of the censorship hypothesis, or must be regarded as an important counter-example? The following could be imposed as a minimum set of conditions for this purpose. Firstly, the naked singularity has to be visible at least for a finite period of time to any far away observer. If only a single null geodesic escaped, it would provide only an instantaneous exposure to the observer by means of a single wave front. In order to yield any observable consequences, a necessary condition is that, a family of future directed non-spacelike geodesics should terminate at the naked singularity in past. Next, this singularity must not be gravitationally weak, creating a mere space-time pathology, but must be a strong curvature singularity in a suitable sense as characterized in Section 5.4. This would ensure that the space-time does not admit any continuous extension through the singularity in any meaningful manner, and hence such a singularity cannot be avoided. The physical effects should then be important near such a strong curvature singularity. Finally, the form of matter should be reasonable in that it must satisfy a suitable energy condition ensuring the positivity of energy, and collapses gravitationally from an initial spacelike hypersurface with a well-defined non-singular initial data. We study here the phenomena of gravitational collapse of a spherical shell of radiation in this context and examine the nature and structure of the resulting naked singularity.

The imploding radiation is described by the Vaidya space-time, given in (v, r, θ, ϕ) coordinates as

$$ds^2 = -\left(1 - \frac{2m(v)}{r}\right) dv^2 + 2dvdr + r^2 d\Omega^2, \tag{6.12}$$

where $d\Omega^2 = d\theta^2 + \sin^2\theta\, d\phi^2$. The radiation collapses at the origin of coordinates, $v = 0, r = 0$. Throughout the discussion here the null coordinate v denotes the advanced time and $m(v)$ is an arbitrary but non-negative increasing function of v. The stress-energy tensor for the radial flux of radiation is

$$T_{ij} = \rho k_i k_j = \frac{1}{4\pi r^2}\frac{dm}{dv} k_i k_j, \tag{6.13}$$

with

$$k_i = -\delta^v{}_i, \quad k_i k^i = 0,$$

which represents the radially inflowing radiation along the world lines $v =$ const. Note that

$$\frac{dm}{dv} \geq 0,$$

implies that the weak energy condition is satisfied. The situation is that of a radially injected radiation flow into an initially flat and empty region, which is focused into a central singularity of growing mass by a distant source (see Fig. 47). The source is turned off at a finite time T when the field settles to the Schwarzschild space-time. The Minkowski space-time for $v < 0, m(v) = 0$ here is joined to a Schwarzschild space-time for $v > T$ with mass $m_0 = m(T)$ by way of the Vaidya metric (6.12). Assuming $m(v)$ to be a linear function, the central singularity $v = 0, r = 0$ was studied by Papapetrou (1985) and Kuroda (1984), who showed that it will be naked and persistent when the collapse is sufficiently slow. The radial null geodesics of this space-time were examined by Papapetrou which clarified the null structure of the space-time for a linear mass function. The particle creation effects associated with such a shell-focusing singularity were studied by Hiscock, Williams and Eardley (1982). They considered the situation when the space-time admits marginally naked singularity, where the Cauchy horizon coincides with the event horizon allowing only an isolated null trajectory to escape to infinity; and studied the particle creation by such a naked singularity. The other cases when the mass function is not linear are discussed by Lake (1986), Rajagopal and Lake (1987) and others. We discuss this situation in detail in Section 6.6.

Our purpose now is to examine the structure and curvature strength of naked singularity arising in the radiation collapse. We would first like to specify here all the families of future directed non-spacelike geodesics

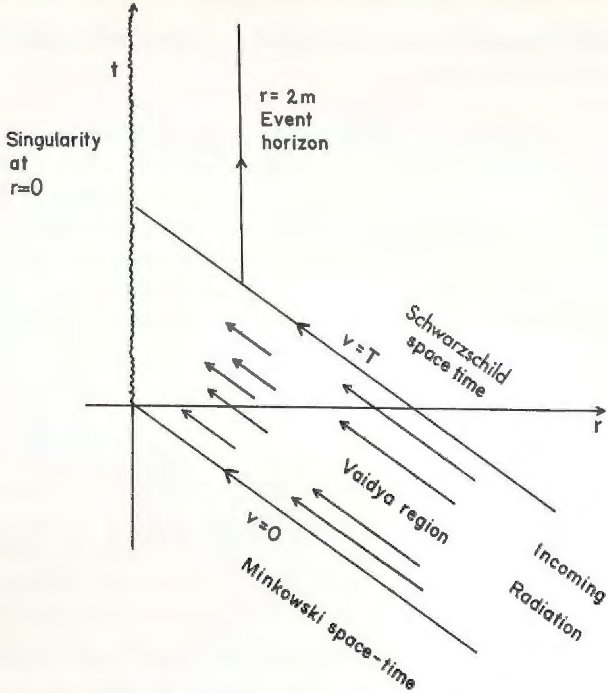

Fig. 47 Collapse of a radiation shell at the center of symmetry. The first wave front arrives at $r = 0$ at $v = 0$ and the final wave front reaches the center at $v = T$. A singularity of growing mass develops at $r = 0$. Here the flat Minkowski space-time for $v < 0$ is joined to the Schwarzschild geometry for $v > T$, through the Vaidya region. The space-time satisfies positivity of energy provided $dm/dv > 0$.

which might possibly terminate at the singularity $v = 0, r = 0$ in the past, thus producing a naked singularity of the space-time. Working out all such families gives a good idea of the nature and structure of the naked singularity as we shall show below, and it also allows one to decide if this is a strong curvature singularity in a stronger sense as discussed in Section 5.4. Next, we examine the curvature growth along such families in the limit of approach to the naked singularity. This allows an assessment of the curvature strength which was done for the case of radial null geodesics by Hollier (1986). Throughout the present section, we choose $m(v)$ to be a linear function,

$$2m(v) = \lambda v,$$

with $\lambda > 0$ (Dwivedi and Joshi, 1989, 1991). This is the Vaidya–Papapetrou space-time describing radiation collapse. To be specific,

$$\begin{aligned} m(v) &= 0 \quad &&\text{for} \quad v < 0, \\ 2m(v) &= \lambda v \quad &&\text{for} \quad 0 < v < T, \\ m(v) &= m_0 \quad &&\text{for} \quad v > T. \end{aligned}$$

Thus, the mass for the final Schwarzschild black hole is m_0 and the causal structure of the space-time depends on the values chosen for the constants m_0, T, and λ. In this case, the Vaidya region (6.12) admits a homothetic Killing vector

$$\xi = v\left(\frac{\partial}{\partial v}\right) + r\left(\frac{\partial}{\partial r}\right),$$

which is given by the Lie derivative,

$$L_\xi g_{ij} = \xi_{i;j} + \xi_{j;i} = 2g_{ij}. \tag{6.14}$$

Let now K^i be tangent to non-spacelike geodesics with $K^i = dx^i/dk$ where k is the affine parameter. Then, $K^i{}_{;j}K^j = 0$ and

$$g_{ij}K^iK^j = B, \tag{6.15}$$

where $B = 0$ for null vectors and $B = \mp 1$ for timelike and spacelike vectors respectively. Along a non-spacelike geodesic,

$$\frac{d}{dk}(\xi^i K_i) \equiv (\xi^i K_i)_{;j}K^j = \xi_{i;j}K^iK^j. \tag{6.16}$$

Then using eqns (6.14) and (6.15) gives $(d/dk)(\xi^i K_i) = B$, which gives on integration

$$\xi^i K_i = Bk + C,$$

where C is a constant. Using the expression for the Killing vector,

$$rK_r + vK_v = Bk + C. \tag{6.17}$$

The components K^θ and K^ϕ of the tangent vector are obtained directly from integration of Lagrange equations,

$$K^\theta = \frac{\ell \cos\beta}{r^2 \sin^2\theta}, \tag{6.18}$$

$$K^\phi = \frac{\ell \sin\beta \cos\phi}{r^2}, \tag{6.19}$$

where ℓ is the impact parameter and β is the isotropy parameter given by $\sin\phi\tan\beta = \cot\theta$. Writing $2m(v) = \lambda v$ and defining a new parameter $X = v/r$, eqn (6.15) gives

$$(1-\lambda X)K_r^2 + 2K_rK_v + \frac{\ell}{r^2} = B. \tag{6.20}$$

Substituting for K_v from eqn (6.17) gives an equation for K_r from the above which can be solved for K_r. Then one could solve for K_v as well. We write this general solution in the following form:

$$K^v = \frac{dv}{dk} = \frac{P(v,r)}{r}, \tag{6.21}$$

$$K^r = \frac{dr}{dk} = \frac{1-\lambda X}{2r}P - \frac{\ell^2}{2rP} + \frac{Br}{2P}. \tag{6.22}$$

Here k is the affine parameter along the geodesics and along radial curves we have $\ell = 0$. The constant B characterizes different classes of geodesics, that is, $B = 0$ for null curves, $B < 0$ for timelike curves, and $B > 0$ for spacelike curves. The function P satisfies the differential equation

$$\frac{dP}{dk} = \frac{P^2}{2r^2}(1-2\lambda X) + \frac{\ell^2}{2r^2} + \frac{B}{2} \tag{6.23}$$

The general solutions for the above equation have been worked out by Dwivedi and Joshi (1989) and are given by

$$P = \frac{(C+Bk) \pm \sqrt{(C+Bk)^2 + (\ell^2 - Br^2)X(2+\lambda X^2 - X)}}{2+\lambda X^2 - X}. \tag{6.24}$$

Here C is a constant and the affine parameter $k = 0$ at $r = 0, v = 0$. From eqns (6.21), (6.22), and the above expression for P we can write

$$\frac{dr}{dv} = \frac{(1+Ak) \mp (1+\lambda X^2 - X)Q}{X(1+Ak) \pm Q)}, \tag{6.25}$$

where $A = B/C$, $L = \ell/C$, and $r = r(X)$ will be used throughout with

$$Q = Q(X) = [(1+Ak)^2 + (L^2 - A/Cr^2)X(2+\lambda X^2 - X)]^{1/2}.$$

The point $r = 0, v = 0$ is a singular point of the above differential equation and is seen to be a naked singularity of the space-time which would be of interest to us.

The basic classification of families of radial as well as non-radial outgoing non-spacelike geodesics terminating at this singularity in the past is best given in terms of the limiting values of X along a singular geodesic in the limit of approach to the singularity. We shall determine here all possible families of singular geodesics. In order to determine the nature of limiting value of X at $r = 0, v = 0$ on singular geodesics, let

$$\lim_{v \to 0, r \to 0} X = \lim_{v \to 0, r \to 0} \frac{v}{r} = X_0,$$

on a singular geodesic. Using eqn (6.25) and the l'Hospital's rule we get

$$X_0 = \lim_{v \to 0, r \to 0} \frac{v}{r} = \lim_{v \to 0, r \to 0} \frac{(dv/d\lambda)}{(dr/d\lambda)} = \frac{X_0[1 \pm Q(X_0)]}{(1 \mp (1 + \lambda X_0^2 - X_0)Q(X_0))}. \tag{6.26}$$

This gives, on simplification,

$$(2 + \lambda X_0^2 - X_0)Q(X_0) = 0. \tag{6.27}$$

Thus, we have either

$$(2 + \lambda X_0^2 - X_0) = 0 \Rightarrow X_0 = a_\pm = \frac{1 \pm \sqrt{1 - 8\lambda}}{2\lambda}, \tag{6.28}$$

or

$$Q(X_0) = 0 \Rightarrow 1 + L^2 X_0(2 + \lambda X_0^2 - X_0) = 0. \tag{6.29}$$

It follows that on singular geodesics the only possible limiting values for X are $a_\pm$ or any one of the roots of the above cubic equation. It is clear from eqn (6.28) that the values $a_\pm$ are possible only if $\lambda \leq 1/8$. It was pointed out by Dwivedi and Joshi (1989) that the singularity $r = 0, v = 0$ becomes a node for families of non-spacelike geodesics provided $\lambda \leq 1/8$ (see Fig. 48). In this case the singularity is naked in the sense that families of future directed non-spacelike geodesics going to infinity terminate at the singularity in the past. A Penrose diagram of this radiation collapse is shown in Fig. 49.

In fact, the five possible values of the tangent indicated above also require this condition on λ. Even though the above gives all the possible values of tangents at which the non-spacelike geodesics can meet the naked singularity in past, all of them need not be actually realized along the trajectories. To see what are the possible values actually realized along singular geodesics, consider the equation for non-spacelike geodesics in the for $r = r(X)$. From the above and $X = v/r$, (we consider positive sign solutions in eqn (6.25)) we can write

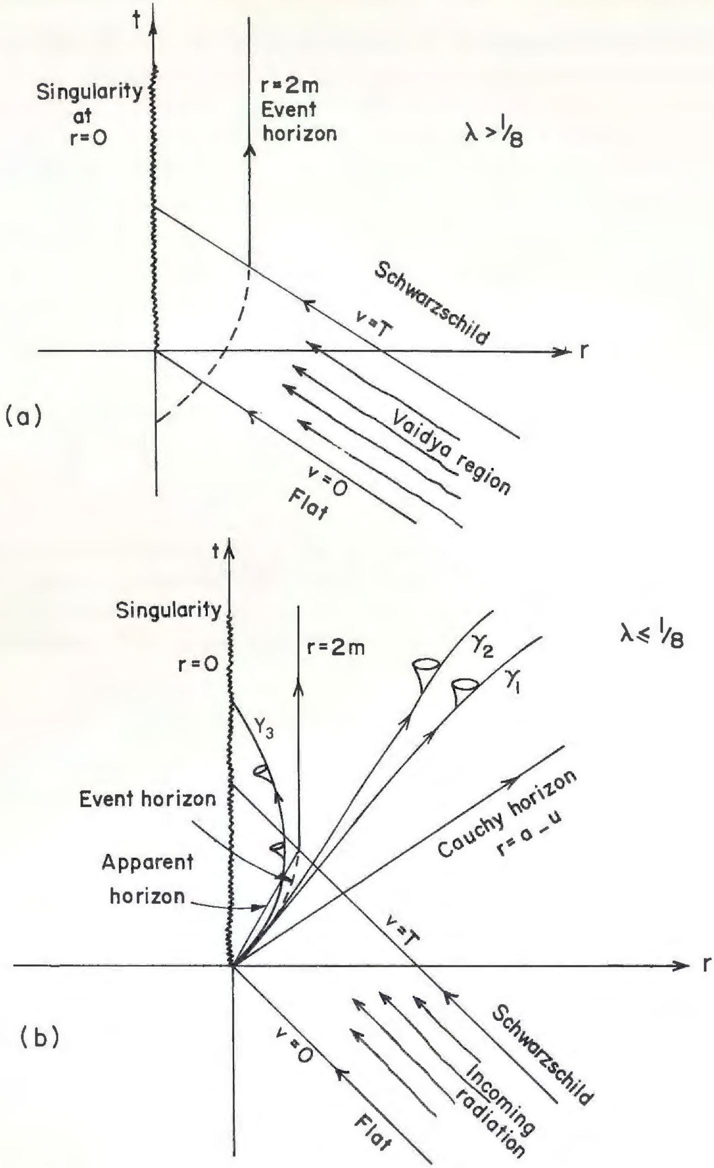

Fig. 48 (a) The event horizon completely covers the singularity at $r = 0$ when $\lambda > 1/8$. (b) For $\lambda \leq 1/8$, a naked singularity forms at the origin. Families of trajectories such as γ_1 and γ_2 escape away to infinity from the singularity with a definite tangent. The non-spacelike curve γ_3, which is emitted after the event horizon, crosses the apparent horizon and falls back to the singularity. The Cauchy horizon is the first ray coming out from the singularity.

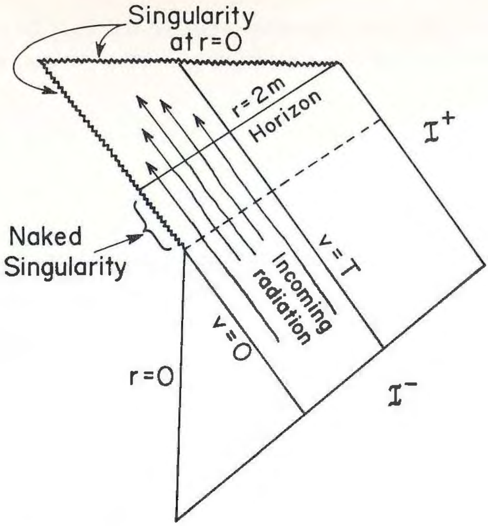

Fig. 49 A Penrose diagram for the naked singularity forming in the radiation collapse.

$$\frac{dr}{dX} = \frac{r[1 + Ak - (1 + \lambda X^2 - X)Q(X)]}{(2 + \lambda X^2 - X)XQ(X)}. \tag{6.30}$$

Integration of the above yields the equation of geodesic curves in (v, r) plane to give $r = r(X)$ for $\lambda \leq 1/8$:

$$\mathrm{D}r = \frac{1}{X-4}\exp\left(\frac{-4}{X-4} + \int \frac{-L^2 + (A/C)r^2}{Q(Q+1+Ak)}dX\right) \quad \text{for } \lambda = 1/8, \tag{6.31}$$

$$\mathrm{D}r = \frac{(X-a_+)^{a_-/(a_+-a_-)}}{(X-a_-)^{a_+/(a_+-a_-)}}\exp\left(\int \frac{-L^2 + (A/C)r^2}{Q(Q+1+Ak)}dX\right) \quad \text{for } \lambda < 1/8. \tag{6.32}$$

where D is a constant which labels different geodesics in the (v, r) plane. For example, $D = 0$, or $D = \infty$ implies $X = v/r =$ const. and gives rise to geodesics which are rectilinear (straight line) in the (v, r) plane. The behaviour of singular geodesics near the singularity is immediate from eqns (6.31) and (6.32). For families of non-rectilinear geodesics (that is, $\mathrm{D} \neq 0, \infty$), terminating at $r = 0$, the allowed values of X are obtained by simply letting $r \to 0$ in eqn (6.32) and finding the corresponding values of X. It follows that either $X = a_+$, or $X = c$, where c is the double root of the eqn (6.29), are the only possible values. To see this note from eqn (6.32), which gives the equation of geodesic curves, that $r = 0$ implies either $X = a_+$ or

$$\exp\left(\int -\frac{dX}{Q(X)}\right) \to 0,$$

which mean $Q(X) \to 0$ for some real value of $X = X_c$. That means X_c should be the root of $Q(X) = 0 \Rightarrow 1 + L^2 X(2 + \lambda X^2 - X) = 0$. This is a third degree algebraic equation and can have three roots; however, since L^2, λ are positive, it can have only two positive and one negative roots. If all the roots are distinct then

$$\int \frac{dX}{Q(X)} \simeq \int \frac{dX}{\sqrt{X - X_c}} \to 0,$$

at any of the distinct roots $X = X_c$ and hence $r \neq 0$. However, if $X = X_c$ is a double root then

$$\int \frac{dX}{Q(X)} \simeq \int \frac{dX}{X_c - X} \to \infty,$$

and therefore $r = 0$ at $X = X_c$. Note that value of the parameter L characterizes the nature of roots of $1 + L^2 X(2 + \lambda X^2 - X) = 0$, and the double root exists if and only if

$$L^2 C^2 \equiv \ell^2 = \ell^2_{\mathrm{crt}} = \frac{27 C^2 \lambda^2}{2(1 - 9\lambda + (1 - 6\lambda)^{\frac{3}{2}})}. \tag{6.33}$$

Since rectilinear geodesics are of interest to us, we specify all singular geodesics that are rectilinear in (v, r) plane. The calculation is straightforward and we find using equations (6.21), (6.22), (6.23), (6.24), and (6.32) that in all there are three rectilinear solutions of geodesic equations. For the radial case when $\ell = 0$,

$$2 + \lambda X^2 - X = 0 \Rightarrow X = \frac{1 \pm \sqrt{1 - 8\lambda}}{2\lambda} = a_\pm, \tag{6.34}$$

$$K^v = 2C_1 k^{(2 - a_\pm)/a_\pm}, \qquad K^r = \frac{2C_1}{a_\pm} k^{(2 - a_\pm)/a_\pm}, \tag{6.35}$$

$$r = C_1 k^{2/a_\pm}, \qquad v = C_1 a_\pm k^{2/a_\pm}, \tag{6.36}$$

where C_1 is a constant. For the non-radial rectilinear geodesic we get

$$2 + 3\lambda X^2 - 2X = 0 \Rightarrow X = \frac{1 + \sqrt{1 - 6\lambda}}{3\lambda} = c, \tag{6.37}$$

$$K^v = \frac{\ell}{r\sqrt{2\lambda c - 1}}, \tag{6.38}$$

$$K^r = \frac{\ell}{cr\sqrt{2\lambda c - 1}}, \tag{6.39}$$

$$r^2 = \frac{2\ell k}{c\sqrt{2\lambda c - 1}}, \qquad X = \frac{v}{r} = c. \tag{6.40}$$

One could easily see that these rectilinear geodesics meet the initial point of the singularity $v = 0$, $r = 0$ provided $a_{\pm}$,c and the tangents are real. This implies that $\lambda \leq 1/8$. We could summarize the results therefore as follows. The three null lines $X = a_+, X = a_-$, and $X = c$ in the (v, r) plane represent the three rectilinear singular geodesics that terminate at the singularity with values of $X = a_+, a_-$, and c respectively. All other families of singular geodesics terminate at the singularity with either $X = a_+$ or $X = c$. Here $X = a_-$ is a single radial geodesic in the (v, r) plane meeting the singularity, which is not tangent to any families of non-spacelike geodesics terminating at the naked singularity. We can therefore classify singular geodesics into three types, depending on the limiting value of X at $r = 0, v = 0$

$$(i) X = a_+, \qquad (ii) X = c, \qquad (iii) X = a_-. \tag{6.41}$$

A point to note is that for $\lambda = 1/8$, $a_+ = a_- = c$. We should mention here that negative sign solutions in eqn (6.24) are not of interest to us as a similar calculation shows that they do not terminate at the singular point $v = 0, r = 0$ at all.

It is worth noting that a geodesic rectilinear in the (v, r) plane need not necessarily be rectilinear in the (v, r, θ, ϕ) plane. For example, the radial rectilinear geodesic $X = a_+$ is rectilinear even in the (v, r, θ, ϕ) plane (since $\phi, \theta =$ const.), while the geodesic $X = c$ which is rectilinear in the (v, r) plane is non-rectilinear in the (v, r, θ, ϕ) plane because $\ell \neq 0$ and $\phi, \theta \neq$ const.

The behaviour of singular geodesics could be better described in terms of the four different regions of the space-time, since singular geodesics do not cross over from one region to other. In region I characterized by

$$X \geq 1/\lambda,$$

all geodesics are ingoing and no geodesics escape. Here $X = 1/\lambda$ is the apparent horizon. Within region II,

$$1/\lambda > X > a_+,$$

the region $(1+\sqrt{1-4\lambda})/2\lambda > X > a_+$ is a free for all region where all geodesics with all value of ℓ escape from the singularity with $X = a_+$. However for $X \geq (1+\sqrt{1-4\lambda})/2\lambda$, family of geodesics with $\ell^2 < (1-\lambda X)/(1+\lambda X^2 - X)^2$ escape while others fall back to $r = 0$.

Region III,

$$a_+ > X > a_-,$$

is separated from region II by the rectilinear radial null geodesic $X = a_+$. Behavior of geodesics in this region could best be understood by tracing the history backwards. Thus all families of geodesics with $\ell < \ell_{\text{crt}}$ all over this region have their past end point at singularity with $X = a_+$. Geodesics with $\ell > \ell_{\text{crt}}$ have their past end point at singularity only within $a_+ > X > c$, while geodesics with $\ell = \ell_{\text{crt}}$ within this region have their past end point at singularity with tangent $X = c$, which is the rectilinear non-radial geodesic. In region IV, no non-spacelike geodesic has an endpoint at the singularity in the past.

Having specified the families of non-spacelike geodesics which emerge from the naked singularity $r = 0, v = 0$, it is our aim now to investigate the structure of the singularity by examining the growth of curvature along different families of non-spacelike geodesics that terminate at the singularity from different directions (that is, $X = a_+$, $X = a_-$, $X = c$). We first examine the strength by considering the following scalar for the Vaidya space-time:

$$\psi = R_{ab}K^aK^b = \frac{\lambda}{r^2}(K^v)^2, \tag{6.42}$$

where R^{ab} is the Ricci curvature tensor. For non-spacelike geodesics the geodesic equations given above imply in general,

$$\begin{aligned}\psi &= \frac{\lambda C^2(1+\mathcal{B}Ck+Q(X))^2}{r^4(2+\lambda X^2-X)^2}\\ &= \frac{4\lambda C^2(1+\mathcal{B}Ck)^2}{r^4(2+\lambda X^2-X)^2} + \frac{(L^2-\mathcal{B}r^2)\lambda X C^2(Q(X)+3+3\mathcal{B}Ck)}{r^4(2+\lambda X^2-X)(Q(X)+1+\mathcal{B}Ck)}.\end{aligned} \tag{6.43}$$

The $\lim_{k\to 0} k^2\psi$ in general can be computed as follows. Multiplying ψ by k^2 and taking the limit as $k \to 0$, using geodesic eqns (6.32), and the above gives

$$\begin{aligned}\lim_{k\to 0} k^2 R_{ab}K^aK^b &= \lim_{k\to 0} k^2\psi\\ &= \lambda C^2\left[\lim_{k\to 0}[1+\mathcal{B}Ck+Q(X)]^2\left(\lim_{k\to 0}\frac{k^2}{r^4(2+\lambda X^2-X)^2}\right)\right].\end{aligned}$$

Since at the singularity $k=0$, $Q(X)$ is positive, $X = X_0$, hence $(1+\mathcal{B}Ck+Q(X))^2$ has a well-defined limit while the quantity in the second bracket has both denominator and numerator approaching zero. We therefore use l'Hospital's rule, geodesic equations, and eqn (6.30) to get

$$\lim_{k\to 0} k^2\psi = \lambda \left(\frac{(1+\sqrt{1+L^2X_o(2+\lambda X_o^2 - X_o)}) \quad X_o}{2+(X_o-2)\sqrt{1+L^2X_o(2+\lambda X_o^2-X_o)}} \right)^2 . \tag{6.44}$$

It is worth noting at this point that for $\infty \geq X_0 \geq -\infty$, $\lim_{k\to 0} k^2\psi > 0$ and therefore the strong curvature condition of Section 5.4 is satisfied along all singular geodesics. Hence it follows that this is a strong curvature singularity in a very powerful sense as characterized in Section 5.4.

In order to examine the behaviour of ψ along families of non-spacelike geodesics near the singularity one must find the dependence of X on k. We get, using eqns (6.21) and (6.22)

$$\frac{dX}{dk} = \frac{1}{r}\left(\frac{dv}{dk} - X\frac{dr}{dk}\right) = \frac{Q(X)}{r^2} \Rightarrow k = \int \frac{r^2(X)dX}{Q(X)}. \tag{6.45}$$

The non-rectilinear geodesics meet the singularity with either $X = a_+$, or $X = c$. We get near the singularity using equations (6.21), (6.22), (6.38), (6.42), and (6.32) for families of geodesics with $X = a_+$ at the singularity,

$$X - a_+ \propto k^n, \qquad n = \sqrt{1-8\lambda}, \qquad r \propto k^{(1-n)/2}, \tag{6.46}$$

$$\psi \simeq \frac{4\lambda}{k^2} + \frac{2\ell^2 D}{k^{(2-n)}}, \tag{6.47}$$

where D is a constant which is negative in region II and positive in region I. Using a similar procedure we calculate the behaviour of ψ along all singular geodesics, rectilinear or otherwise, and the results can be described as follows.

We discuss all the three classes of singular geodesics in terms of the limiting values of X at the naked singularity, namely, $X = a_+, c, a_-$. As we see below, the strong curvature condition is satisfied in all the cases, that is, along all the families of singular non-spacelike geodesics.

Firstly, for the class $X = a_+$, for the radial or rectilinear singular geodesics the behaviour of ψ is given by

$$\psi = 4\lambda/k^2,$$

and

$$\lim_{k\to 0} k^2\psi = 4\lambda.$$

For all other geodesics ($\ell \neq 0$), we have again

$$\psi \simeq (4\lambda/k^2) + (2\ell^2 D/k^{(2-n)}),$$

and again

$$\lim_{k\to 0} k^2\psi = 4\lambda.$$

Next, for $X = c$ value, the behaviour of the curvature scalar ψ for the rectilinear non-radial trajectories is given by

$$\psi = \lambda c^2/4k^2,$$

and the limit above for $k^2\psi$ works out to be $\lambda c^2/4$. For non-radial trajectories in this class, given by $\ell = \ell_{crt}$, we have $\psi \simeq \lambda c^2/4k^2$ which gives

$$\lim_{k\to 0} k^2\psi = \frac{\lambda c^2}{4}.$$

Finally, for the rectilinear radial geodesic with $X = a_-$, one has

$$\psi = \frac{4\lambda}{k^2},$$

and

$$\lim_{k\to 0} k^2\psi = 4\lambda.$$

Thus the strong curvature condition is again satisfied.

The important point to note is that the limiting value of $k^2\psi$ for all non-spacelike geodesics that terminate at the singularity is always positive, even though its actual value is path dependent. Away from the singularity the growth of $k^2\psi$ for non-radial geodesics shows deviation. This is due to the contribution of non-radial terms in eqn (6.43). In region I of the spacetime non-radial terms add to the growth while in region II they check the growth. This results in a somewhat unusual peak in the value of $k^2\psi$ in region III and it can be seen that the peak increases as the value of ℓ gets nearer to ℓ_{crt}. For ℓ further than ℓ_{crt} the growth aligns itself more and more as $1/k^2$ and for $\ell = 0$, or $\ell = \infty$, becomes as $1/k^2$ (see Fig. 50).

Having seen that the Vaidya–Papapetrou naked singularity is a strong curvature singularity, one would expect it to be a scalar polynomial singularity as well. This is another notion matching well with our concept of singular behaviour familiar from 'big-bang' cosmologies (divergence of the

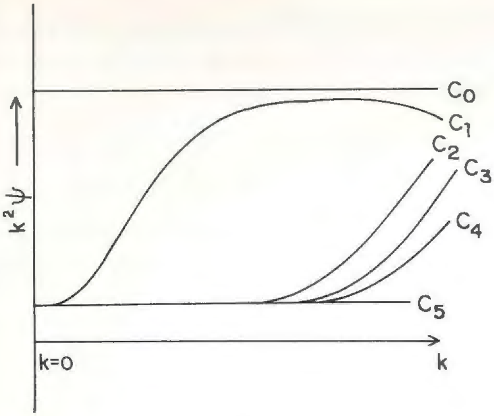

Fig. 50 The curvature growth along different non-spacelike trajectories meeting the naked singularity in the past for the radiation collapse models.

Ricci scalar) and the Schwarzschild models (divergence of the Kretschmann scalar $R^{ijkl}R_{ijkl}$). In fact, all known cases of strong curvature singularities in Einstein's equations are scalar polynomial singularities as well. Any such singularity will be an end point of at least one curve on which a scalar polynomial in the metric and the Riemann tensor takes unboundedly large values.

We therefore examine now the behaviour of the Kretschmann scalar near the naked singularity forming in the Vaidya–Papapetrou models. The Kretschmann scalar is given by

$$\alpha = R^{ijkl}R_{ijkl} = \frac{48M(v)^2}{r^6} = \frac{12\lambda^2 X^2}{r^6}, \tag{6.48}$$

along the families of the non-spacelike geodesics joining the singularity. It is verified that this scalar always diverges, thus establishing the scalar polynomial singularity as expected. However, an interesting directional behaviour is revealed by the singularity as far as the scalar α is concerned, which was not the case for the scalar ψ examined above. Using the relation $r = r(X)$ given by eqns (6.32), (6.36), and (6.40), and $X = X(k)$ from eqn (6.46), calculations are straightforward and the behaviour of the Kretschman scalar α along different families near the singularity can be described as follows.

Again, we discuss the behaviour of α near the naked singularity in terms of the limiting value of X. For the class given by $X = a_+$,

$$\alpha = \frac{12\lambda^2 a_+^2}{C_1^4 k^{8/a_+}}.$$

Then, $\lim_{k\to 0} k^2\alpha = 0$ and also, $\lim_{\lambda\to 0} \alpha|_{k=k_0} = 12/C_1^4$. It thus turns out that the divergence of α is λ dependent.

Next, for the case $X = c$,

$$\alpha = \frac{12\lambda^2 c^4(2\lambda c - 1)}{\ell_{crt}^2 k^2},$$

and the limit $\lim_{k\to 0} k^2\alpha$ works out to be $12\lambda^2 c^4(2\lambda c - 1)/\ell_{crt}^2$. In this case, $\lim_{\lambda\to 0} \alpha|_{k=k_0} = 0$ and the divergence of α turns out to be λ independent.

Finally, for the case $X = a_-$,

$$\alpha = \frac{12\lambda^2 a_-^2}{C_1^4 k^{8/a_-}}.$$

Then both the limits discussed above turn out to be infinite and the divergence of α is again λ dependent.

Unlike the scalar ψ whose behaviour was indifferent to the direction of approach to singularity, the Kretschmann scalar not only shows directional dependence but also a dependence on the parameter λ which characterizes the rate at which the null dust is imploding and is bound by the condition $\lambda \leq 1/8$. Another point in the behaviour of α is the case when the null dust is imploding towards the center very slowly, that is, for very small values of λ. Although α always diverges sufficiently close to the singularity, it is interesting to note the growth of α in the limit $\lambda \to 0$ as described above, where the evaluation of the limit is at an arbitrary but fixed value of the affine parameter, $k = k_0$ along the trajectory. Of course for any arbitrarily small but fixed value of λ, in the limit as the affine parameter k tending to zero, the scalar α always diverges as shown earlier.

It is clear from the above consideration that the naked singularity exhibits a directional behaviour as far as the Kretschmann scalar is concerned along different families. Such a situation has been referred to as a 'directional singularity', where the singularity strength varies with direction. We point out that various authors have discussed possible interpretations for this situation. In fact, Ellis and King (1974) discussed such a directional property within the cosmological scenario of Friedman models where the strength depends on the direction along which the geodesic enters the singularity. We have now shown that a similar property is exhibited by the naked singularity resulting from gravitational collapse.

To summarize the above results, not just isolated trajectories but entire families of non-spacelike geodesics escape from the naked singularity at the origin outside the event horizon and inside the Cauchy horizon $X = a_-$. As a result, the naked singularity is visible to a far away observer for an infinite period of time once he or she has received the first ray from the singularity. The structure and details of these families are specified above. Further, the analysis of curvature growth along the singular geodesics shows that this is a strong curvature naked singularity in a very powerful sense.

6.6 Self-similar gravitational collapse

The existence of a strong curvature naked singularity is established in the previous section for the radiation collapse, which could be regarded as a serious example of a naked singularity in the sense described there. However, these models could be considered deficient in one respect, namely that the equation of state used may not be considered realistic in somewhat earlier phases of the collapse. While it is reasonable to assume that during the late stages of collapse almost all the mass of the star is converted into radiation (see for example, Harrison, Thorne, Wakano and Wheeler, 1965) and that a collapsing radiation sphere might be a good model for such late epochs, it would be clearly advantageous to have a situation which considers matter with a reasonable equation of state which includes non-zero pressures as well which could play an important role for the later stages of collapse. The point is, for a reasonable equation of state the possibility remains that the pressure gradients might prevent the occurrence of a naked singularity or could reduce the strength of the name. Apart from this, another important generalization one would like to include is to allow for inhomogeneous distribution of matter. The reason is, the exactly homogeneous Oppenheimer–Snyder dust ball collapse described in Section 6.2 is a highly idealized model and a realistic collapse would probably start from a quite inhomogeneous initial distribution of matter with higher densities near the center.

The purpose of this section is to generalize the conclusions on naked singularity of the previous section, namely, the existence, the termination of non-spacelike geodesic families, and the strength of the singularity, for the general class of self-similar space-times with a physically realistic matter distribution. Specifically, we study here the scenario for the gravitational collapse of a perfect fluid obeying an adiabatic equation of state in a general spherically symmetric self-similar space-time and examine the structure and formation of naked singularity. In fact, the Vaidya solutions discussed in the previous section are self-similar when the mass function is linear. In that sense the results here provide a direct generalization of these earlier results for the general class of self-similar space-times.

A self-similar space-time is characterized by the existence of a homothetic Killing vector field (Cahill and Taub, 1971). A spherically symmetric space-time is self-similar if it admits a radial area coordinate r and an orthogonal time coordinate t such that for the metric components g_{tt} and g_{rr}

we have

$$g_{tt}(ct, cr) = g_{tt}(t, r),$$

$$g_{rr}(ct, cr) = g_{rr}(t, r),$$

for all $c > 0$. Thus, along the integral curves of the Killing vector field all points are similar. For a self-similar space-time, the Einstein equations reduce to ordinary differential equations which we analyse here. The main application of the self-similar space-times so far has been in the cosmological context (see for example, Eardley, 1974; Carr and Hawking, 1974). However, their use to describe gravitational collapse is not ruled out and what is required is a matching of a self-similar interior to a Schwarzschild exterior space-time. It was shown by Ori and Piran (1990) that such a matching can be done as smoothly as desired and that if the matching is sufficiently far from the center, the central region evolves in a self-similar manner without being affected by the matching.

A spherically symmetric space-time in comoving coordinates is given by

$$ds^2 = -e^{2\nu(t,r)}dt^2 + e^{2\psi(t,r)}dr^2 + r^2S^2(t,r)(d\theta^2 + \sin^2\theta\, d\phi^2). \tag{6.49}$$

Self-similarity implies that all variables of physical interest may be expressed in terms of the similarity parameter $X = t/r$. Therefore ν, ψ and S are functions of X only. The pressure and energy density in the comoving coordinates are $(u^a = e^{-\nu}\delta^a_t)$

$$P = \frac{p(X)}{8\pi r^2}, \qquad \rho = \frac{\eta(X)}{8\pi r^2}. \tag{6.50}$$

The field equations for a self-similar collapse of a spherically symmetric perfect fluid are (see for example, Cahill and Taub, 1971),

$$\begin{aligned} G^t_t = \frac{-1}{S^2} + \frac{2e^{-2\psi}}{S}\left(X^2\ddot{S} - X^2\dot{S}\dot{\psi} + XS\dot{\psi} + \frac{(S - X\dot{S})^2}{2S}\right) \\ - \frac{2e^{-2\nu}}{S}\left(\dot{S}\dot{\psi} + \frac{\dot{S}^2}{2S}\right) = -\eta, \end{aligned} \tag{6.51}$$

$$\begin{aligned} G^r_r = \frac{-1}{S^2} - \frac{2e^{-2\nu}}{S}\left(\ddot{S} - \dot{S}\dot{\nu} + \frac{\dot{S}^2}{2S}\right) \\ + \frac{2e^{-2\psi}}{S}\left[-SX\dot{\nu} + X^2\dot{S}\dot{\nu} + \frac{(S - X\dot{S})^2}{2S}\right] = p, \end{aligned} \tag{6.52}$$

$$G_r^t = \ddot{S} - \dot{S}\dot{\nu} - \dot{S}\dot{\psi} + \frac{S\dot{\psi}}{X} = 0. \tag{6.53}$$

The conservation equation is given by

$$T^{ij}{}_{;j} = 0 \Rightarrow \dot{p} - \frac{2p}{X} = -(\eta + P)\dot{\nu},$$

and we have,

$$\dot{\eta} = -(\eta + p)\left(\dot{\psi} + \frac{2\dot{S}}{S}\right), \tag{6.54}$$

where the dot denotes differentiation with respect to the similarity parameter X. Integrating the above and eliminating $\ddot{S}$ from eqns (6.51) and (6.52) we get

$$e^{2\psi} = \alpha\eta^{\frac{-2}{a+1}}S^{-4}, \tag{6.55}$$

$$e^{2\nu} = \gamma(\eta X^2)^{\frac{-2a}{a+1}}, \tag{6.56}$$

$$\dot{V}(X) = Xe^{2\nu}[(\eta + p)e^{2\psi} - 2] = Xe^{2\nu}[H - 2], \tag{6.57}$$

$$\left(\frac{\dot{S}}{S}\right)^2 V + \left(\frac{\dot{S}}{S}\right)(\dot{V} + 2Xe^{2\nu}) + e^{2\nu+2\psi}\left(-\eta - e^{-2\psi} + \frac{1}{S^2}\right) = 0. \tag{6.58}$$

The quantities V and H here are defined by

$$V(X) \equiv e^{2\psi} - X^2e^{2\nu}, \quad H \equiv (\eta + p)e^{2\psi}.$$

While integrating eqn (6.54) we have assumed that the collapsing fluid is obeying an adiabatic equation of state

$$p(X) = a\eta(X),$$

where a is a constant in the range $0 \leq a \leq 1$. Equations(6.57) and (6.58) are valid in general. As pointed out by Cahill and Taub (1971) and Bicknell and Henriksen (1978), the equation of state $p = a\rho$ is compatible with self-similarity and if the initial values of density, first derivatives, and the gravitational potential satisfy the field equations given above, the system will evolve in a self-similar manner. The special case $a = 0$ here describes dust and $a = 1/3$ gives the equation of state for radiation.

It is clear from eqn (6.50) that the point $t = 0$, $r = 0$ is a singularity for all self-similar solutions where the density necessarily diverges. The value of density is infinite at this point and such a divergence is also observed when we approach this singularity on any line of self-similarity $X = X_0$. This leads to the divergence of curvature scalars such as $R^j{}_iR^i{}_j$ and also of the Ricci scalar $R = \rho - 3p$ if $a \neq 1/3$. We would now like to examine when this singularity could be possibly naked, and if so whether families of non-spacelike geodesics would terminate at the same in the past. The next question then would be the determination of curvature strength of this naked singularity in order to decide on its seriousness and physical relevance. For this purpose, we first need to specify the families of non-spacelike geodesics for self-similar space-times. This is possible using the existence of a homothetic Killing vector field for such a space-time.

Let K^i be the tangent to geodesics with

$$K^i K_i = \mathcal{B}. \tag{6.59}$$

The constant $\mathcal{B}$ characterizes different classes of geodesics namely, $\mathcal{B} = 0, \mathcal{B} < 0, \mathcal{B} > 0$ corresponds to null, timelike, and spacelike geodesics respectively. For the self similar metric given by eqn (6.49) the Lagrange equations immediately give K^θ and K^ϕ:

$$K^\theta = \frac{\ell \cos\phi \sin\beta}{r^2 S^2}, \qquad K^\phi = \frac{\ell \cos\beta}{r^2 S^2 \sin^2\theta}, \tag{6.60}$$

where ℓ is the impact parameter and β is the isotropy parameter. In order to calculate K^t and K^r, note that the homothetic killing vector

$$\xi^a = r\left(\frac{\partial}{\partial r}\right) + t\left(\frac{\partial}{\partial t}\right),$$

admitted by the self-similar space-time satisfies

$$(K^i \xi_i)_{;j} K^j = 2K^i \xi_{i;j} K^j = 2\mathcal{B} \Rightarrow K^i \xi_i = C + \mathcal{B}k, \tag{6.61}$$

where C is an integration constant and k is the affine parameter. From the above algebraic equation and the fact that $K^i K_i = \mathcal{B}$ we get, after using expressions for K^θ and K^ϕ,

$$re^{2\psi} K^r - te^{2\nu} K^t = C + \mathcal{B}k, \qquad e^{2\psi}(K^r)^2 - e^{2\nu}(K^t)^2 + \frac{\ell^2}{r^2} = \mathcal{B}. \tag{6.62}$$

Solving the above equations we get

$$K^t = \frac{C(X \pm e^{2\psi} Q)(1 + BCk)}{r(e^{2\psi} - e^{2\nu} X^2)}, \tag{6.63}$$

$$K^r = \frac{C(1 \pm X e^{2\nu} Q)(1 + BCk)}{r(e^{2\psi} - e^{2\nu} X^2)}, \tag{6.64}$$

where

$$Q = Q(X) = \sqrt{e^{-2\psi-2\nu} + \frac{(L^2 - Br^2)(e^{2\psi} - X^2e^{2\nu})e^{-2\psi-2\nu}}{S^2(1 + BCk)^2}}, \tag{6.65}$$

where we have put $B = \mathcal{B}/C^2$ and $L^2 = \ell^2/C^2$. The quantity Q is always taken to be positive and the positive and negative signs represent two sheets of solutions. The positive sign solutions describe the outgoing trajectories while negative ones represent the ingoing solutions reaching the singularity. To show this we first consider negative sign solutions and show that these do not connect to the point $r = 0$ for positive values of X (that is, future outgoing). From equations for dr/dk and dt/dk we get

$$\frac{dt}{dr} = \frac{X \pm e^{2\psi}Q}{1 \pm Xe^{2\nu}Q}, \tag{6.66}$$

$$\frac{dX}{dr} = \frac{\pm(e^{2\psi} - X^2e^{2\nu})Q}{r(1 \pm Xe^{2\nu}Q)}. \tag{6.67}$$

After some rearrangements and simplification we get

$$\begin{aligned}\frac{dt}{dr} &= \frac{(X \pm e^{2\psi}Q)(1 \mp Xe^{2\nu}Q)}{(1 \pm Xe^{2\nu}Q)(1 \mp Xe^{2\nu}Q)} \\ &= \frac{\left(-X(\ell^2 - Br^2) \pm QS^2(C + Bk)^2\right)e^{2\psi}}{S^2(C + Bk)^2 - X^2e^{2\nu}(\ell^2 - Br^2)}.\end{aligned} \tag{6.68}$$

Similarly we get

$$\frac{dX}{dr} = \frac{S^2(C + Bk)^2Qe^{2\psi}(-Xe^{2\nu} \pm 1)}{rS^2(C + Bk)^2 - X^2e^{2\nu}(\ell^2 - Br^2)}. \tag{6.69}$$

Hence for solutions with $(-)$ sign we get from above by integrating

$$r = \exp\left(-\int \frac{S^2(C + Bk)^2 - X^2e^{2\nu}(\ell^2 - Br^2)dX}{S^2(C + Bk)^2Qe^{2\psi}(Xe^{2\nu} + 1)}\right). \tag{6.70}$$

From the equation for dt/dr it follows in case of negative sign solutions that for outgoing geodesics(that is, dt/dr, and X is positive at the singularity), $S^2(C + Bk)^2 - X^2e^{2\nu}(\ell^2 - Br^2)$ should be negative. Hence the integrand inside the above equation are always positive, therefore the geodesics will never meet the singular point $r = 0$ for an outgoing ray. Further, note that in the case where curves are meeting the singularity with a negative value of X, dt/dr is negative and hence negative sign solutions are always ingoing as far as the singularity is concerned. Similarly, positive sign solutions are outgoing and terminate at the singularity with positive value of X.

To proceed further, we choose the function Q to be positive throughout and $\pm$ signs represent outgoing or ingoing solutions. From the above equations we get

$$\frac{dt}{dr} = \frac{X \pm e^{2\psi}Q}{1 \pm Xe^{2\nu}Q}. \tag{6.71}$$

The point $t = 0, r = 0$ is a singular point of the above differential equation. The nature of the limiting value of similarity parameter $X = t/r$ plays an important role in the analysis of non-spacelike curves that terminate at the singularity and reveals the exact nature of the singularity. Using eqn (6.71) and l'Hospital's rule we get

$$X_0 = \lim_{t\to 0, r\to 0} \frac{t}{r} = \lim_{t\to 0, r\to 0} \frac{dt}{dr} = \frac{X_0 \pm e^{2\psi(X_0)}Q(X_0)}{1 \pm X_0 e^{2\nu(X_0)}Q(X_0)}$$

$$\Rightarrow \mathcal{F}(X_0) \equiv V(X_0)Q(X_0) = 0, \tag{6.72}$$

Thus we see that either

$$V(X_0) \equiv e^{2\psi(X_0)} - X_0^2 e^{2\nu(X_0)} = 0, \tag{6.73}$$

or $Q(X_o) = 0$, which implies

$$L^2 V(X_0) = -S^2(X_o). \tag{6.74}$$

If $\mathcal{F}(X_o) = 0$ does not have any real roots then geodesics clearly do not terminate at the singularity with a definite tangent, that is, the singularity is either a focus or center. In case $\mathcal{F}(X_0) = 0$ does have real roots then the singularity could either be a node or a col. In fact it follows from the geodesic equations (6.63) and (6.64) that straight lines in the (t, r) plane given by $X = t/r = X_0$, where X_0 is either any root of $V(X_0) = 0$ or any double root of $Q(X_0) = 0$, represent in (t, r) plane rectilinear radial and non-radial null geodesics respectively. Therefore if $V(X) = 0$ has at least one real positive root or $Q(X_o)$ has a double root, then the singularity is naked. In this case, when $V(X_0) = 0$ has a positive real root, a single radial null geodesic would escape from the singularity. As we see from eqn (6.72), the singularity could be naked even if $V(X_0) = 0$ does not have real positive root. This is the situation when $V(X) < 0$ and a double positive root of $Q(X) = 0$ exists. Existence of the positive real roots of $\mathcal{F}(X) = 0$ is therefore a necessary and sufficient condition that the singularity would be naked and at least one single null geodesic in the (t, r) plane would escape from the singularity. However, even in such a case other non-spacelike geodesics may or may not terminate at the singularity. If only a single null geodesic escapes, it amounts to a single wave front being emitted from the singularity and hence the singularity would appear to be naked only instantaneously to a distant observer. If the naked singularity is to be seen for a finite period of time, a family of integral curves (geodesics) must escape from the singularity.

At this juncture it would be relevant to note the case of the naked singularity arising due to gravitational collapse of null dust in the Vaidya space-times discussed in the previous section. The origin of the coordinates there is a naked singularity which is a node and an entire family of non-spacelike geodesics escape, exposing the singularity to a distant observer for an infinite time. In order to find whether a family of null or timelike geodesics would terminate at the singularity ($t = 0, r = 0$) in the present case, one must analyse the structure of this singularity. We have shown here that the singularity could become a node only if $\mathcal{F}(X) = 0$ has a real root X_o which gives a direction tangent to the integral curves at the singularity. It is of course possible that $X = X_o$ may not be realized along the integral curves, but a single null geodesic $X = X_o$ is escaping out as stated above. To examine this issue we consider the equation of geodesics $r = r(X)$ in (r, X) plane. We shall restrict to positive sign solutions which represent outgoing geodesics as seen above. Using eqns (6.63) and (6.64) we obtain

$$\frac{dX}{dr} = \frac{V(X)Q(X)}{r(1 + Xe^{2\nu}Q)}. \tag{6.75}$$

Integration of the above yields the equation of non-spacelike geodesics (integral curves) which can be written as

$$r = D\exp\left(\int \frac{(1 + Xe^{2\nu}Q)}{V(X)Q(X)}dX\right). \tag{6.76}$$

Here D is a constant that labels different integral curves. We have already established the fact that if the singularity is to be naked, $\mathcal{F}(X) = 0$ must have at least one real positive root X_o. Hence we first consider the case when $V(X) = 0$ has one simple real positive root. Using eqn (6.57) we write near the singularity

$$V(X) = (X - X_o)X_o e^{2\nu(X_o)}(H(X_o) - 2),$$

and use the fact that Q is positive to integrate eqn (6.76) near the singularity. This gives

$$r = D(X - X_o)^{\frac{2}{H_o - 2}}, \tag{6.77}$$

where $H_o = H(X_o)$. When $H_o > 2$ it is seen that an infinity of integral curves will meet the singularity in the past with tangent $X = X_o$; different curves being characterized by different values of the constant D. Hence the singularity $(0, X_o)$ in (r, X) plane is a node (see for example, Nemytskii and Stepanov, 1960) and it corresponds to $r = 0, t = 0$ being a node in the (t, r) plane. It follows that this singularity is at least locally naked from which an infinity of non-spacelike curves are ejected. In the case $H_o < 2$ the singularity $(0, X_o)$ in (r, X) plane is a col and therefore the behaviour of geodesics in (t, r) plane depends on the value of $2/(2-H_o)$. This is because

$$r(X - X_o)^{2/(2-H_o)} = D$$

in the (r, X) plane and

$$Dr^{H_o/(2-H_o)} = (t - X_o r)^{(2/2-H_o)},$$

in the (t, r) plane. If $H_o > 0$ the singularity would be a node in the (r, t) plane. However, if $H_o < 0$ the singularity in the (r, t) plane is a col where integral curves move away from the singularity and never terminate there. It should also be noted that in the case $H_o = \infty$ or $H_o = 0$ the curves do not terminate at the singularity. Therefore we could deduce that the integral curves would terminate at the singularity, which would be a node in the (t, r) plane as long as

$$\infty > H_o = H(X_o) = (\eta + p)e^{2\psi} > 0. \qquad (6.78)$$

It is seen that eqn (6.78) will be satisfied provided the weak energy condition (see Section 5.2) holds and further that the energy density as measured by any timelike observer is positive in the collapsing region near the singularity.

The results could be summarized as follows. If in a self-similar spacetime a single null radial geodesic escapes the singularity, then an entire family of non-spacelike geodesics would also escape (in the sense of the singularity being locally naked) provided the positivity of energy density is satisfied in the above sense. Such a singularity is a node at the origin. It also follows that no families of non-spacelike geodesics would escape the singularity, even though a single null trajectory might, if the weak energy condition is violated.

In order to examine when such a locally naked singularity will be globally naked, so that it is exposed to a distant observer for a finite period of time, we consider the case when $\mathcal{F}(X) = 0$ has two real simple positive roots X_o and X_1. Suppose $X_o > X_1$, then we have from eqn (6.76)

$$r = D \exp\left(\int \frac{(1 + X e^{2\nu} Q)}{(X - X_o) f(X)} dX - \int \frac{(1 + X e^{2\nu} Q)}{(X - X_1) f(X)} dX \right), \qquad (6.79)$$

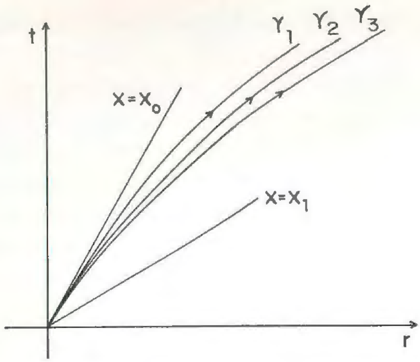

Fig. 51 An open nodal region forming in the self-similar gravitational collapse. Infinitely many non-spacelike curves escape to infinity between the roots $X = X_0$ and $X = X_1$.

where

$$f(X) = \frac{(X_o - X_1)\mathcal{F}(X)}{(X - X_o)(X - X_1)}. \tag{6.80}$$

The function $f(X)$ does not change sign between the interval X_o and X_1 and has the same sign at X_o and X_1. Therefore all the integral curves would terminate at the singularity at one of the roots $X = X_o$ or $X = X_1$. The same trajectories will reach the infinity $r = \infty$ at the other value of the root. Thus the singularity would be globally naked and an infinity of curves would escape from the singularity to reach any distant observer. The region between $X = X_o, X = X_1$ is therefore an open nodal region (Fig. 51).

The conclusions are the same if $\mathcal{F}(X) = 0$ has more than two simple roots. In fact if $\mathcal{F}(X) = 0$ has n simple positive roots $X = X_o, \ldots, X_n$, the directions $X = X_n$ would be alternatively nodes and cols in the (r, X) plane. Another interesting point that emerges from the above consideration is the possibility of occurrence of closed nodal regions. Consider three singular points in the (r, X) plane, namely P_{n-1}, P_n and P_{n+1}, such that P_n is a node and integral curves meet the singularity in the (t, r) plane with tangent $X = X_n^o$, P_{n-1} is col and in the (t, r) plane there is col with separators at the origin, and P_{n+1} is a col and in the (t, r) plane $X = X_{n+1}^o$ is tangent to the integral curves at the singularity. Therefore the region between $X = X_n^o$ and $X = X_{n+1}^o$ in the (t, r) plane could be a closed nodal region where curves emerge with integral curves forming a closed loop. The region between $X = X_n^o$ and $X = X_{n-1}^o$ in the (t, r) plane is an open nodal region where the integral curves emerge from the singularity and escape to infinity. We should, however, note that dt/dr does not vanish for outgoing curves as seen from eqn (6.66) and hence in self-similar space-times a closed nodal region could not exist.

It should be remarked that the existence of real positive roots for $V(X) = 0$ provides a sufficient condition for the occurrence of naked singularity at the origin. It is relevant to ask when this will be realized in terms of the parameters in self-similar field equations. An answer to that can be obtained by some further analysis of the self-similar field equations given above. In eqns (6.55) and (6.56), α and γ are constants of integration and could be set to unity by a suitable scale transformation. Then, by putting

$$y = X^{\beta}, \quad U^2 = \frac{e^{-2\nu+2\psi}}{X^2} = \frac{1}{1 - Ve^{-2\psi}} = \eta^{-2\beta} y^{-2} S^{-4},$$

with $b = 1 + a$, and $\beta = (1 - a/1 + a)$, and using eqns (6.55) and (6.66), eqns (6.57) and (6.58) can be expressed as

$$\beta y \frac{\eta'}{\eta} = \frac{1}{U^2 - a}\left(2a - 2b\beta y U^2 \frac{S'}{S} - \frac{1}{2} b^2 \eta^{\beta} y^2 U^2\right), \tag{6.81}$$

$$\begin{aligned}\left(\frac{S'}{S}\right)^2 \beta^2 y^2 (U^2 - 1) + \beta y \left(\frac{S'}{S}\right)(b\eta^{\beta} y^2 U^2 + 2) \\ - \left(\eta^{\beta} y^2 U^2 (1 - \frac{1}{\eta S^2}) + 1\right) = 0,\end{aligned} \tag{6.82}$$

where the dash denotes differentiation with respect to y. The quantities U and H are scale invariants, that is, invariant under a transformation of the type $t \to ct, r \to dr$ and hence are of physical interest. In fact U represents the velocity of the fluid relative to X = const. hypersurfaces (Bicknell and Henriksen, 1978) while H is related to energy. Solutions of the above equations which allow

$$U(X) = U(y) = 1,$$

for some real positive value of $y = y_0 = X_0^{\beta}$ are of special interest to us. It will be seen that this corresponds to a situation of the space-time containing a naked singularity at the origin of coordinates where non-spacelike geodesics terminate in the past. It should be noted that the point X_p at which $U^2 = a$ is a sonic point where the fluid velocity relative to the similarity lines (X = const.) is equal to the speed of sound. Therefore in all solutions of the field equations which allow the existence of the sonic point, there would also exist at least one point at which $U = 1$. Actually solutions of the field equations which allow $U = 1$ at some real value of X are an initial value problem as the field equations do not have $U = 1$

as a singular point of the differential equations. Our aim therefore is only to determine what the values of different parameters could be for solutions that allow $U = 1$ for some real positive value of X. Hence we analyse now the solutions of above differential equations near the point $y = y_o = X_0^\beta$, with the condition such that

$$U(y_o) = (\eta_o)^{-\beta} y_o^{-2} S_o^{-4} = (\eta_o X_o)^{-2\beta} S_o^{-4} = 1,$$

where $S_o = S(y_o)$ and $\eta_o = \eta(y_o)$. We write

$$\eta(X) = \eta_o + \eta_o \sum_{n=1}^{\infty} \eta_n (y - y_o)^n, \quad S(X) = S_o + S_o \sum_{n=1}^{\infty} S_n (y - y_o)^n. \tag{6.83}$$

By requiring that the above satisfy the differential equation for η and S we get the values of coefficients S_n and η_n in terms of the initial values of η_o and η_1 to obtain

$$S_o^4 = \frac{1}{\eta_o^{2\beta} y_o^2}, \tag{6.84}$$

$$S_1 = \frac{1}{\beta y_o (b\eta^\beta y_o^2 + 2)} \left(\eta^\beta y_o \left(1 - \frac{1}{\eta_o S_o^2}\right) + 1 \right), \tag{6.85}$$

$$\eta_1 = \frac{1}{2\beta y_o (1 - a)} \left(4a - 4\beta y_o S_1 - b^2 \eta_o^\beta y_o^2\right). \tag{6.86}$$

Eliminating S_o and S_1 from the above equations and writing $z = b\eta_o^{\beta/2} y_o$ we get, after some simplification,

$$z^4 + n(m-2)z^3 + 2(3-a)z^2 + 2bnmz + 4b(1-a) = 0, \tag{6.87}$$

where,

$$n = \frac{2\eta_o^{\frac{\beta-2}{2}}}{b}, \quad m = \frac{\beta(1-a)\eta_1}{\eta_o^{\beta-1}}.$$

The above is a fourth order algebraic equation and can have at the most four real roots depending upon the values of the parameters η_0, η_1. Thus, the condition that $V(X) = 0$ has real positive roots is characterized in terms of eqn (6.87) in the sense that the parameters n and m (that is, η_0 and η_1) must be such that the above forth order equation has real positive roots.

We may note here that the issue of global existence of solutions to self-similar Einstein field equations, with the equation of state discussed here and with a physical initial data, has not been addressed so far on a rigorous mathematical level. However, the numerical and qualitative analysis given by Ori and Piran (1990), who treated the issue of sonic point in detail, indicate good reasons to believe that global solutions exist in the self-similar case. One could regard their study as a good motivation to examine the local behaviour of such solutions in the vicinity of naked singularity. In any case, this local behaviour, and in particular the orbits of non-spacelike geodesics that leave the singularity, does not depend on the global question, that is, whether a particular self-similar solution can or cannot be matched to reasonable initial data.

We now examine the curvature strength of this naked singularity forming at $t = 0, r = 0$. Because, even though a naked singularity may form during gravitational collapse, if it is not a strong curvature singularity, it may not be considered a serious counter-example to the cosmic censorship hypothesis. As discussed earlier, a sufficient condition to ensure a strong curvature singularity is given as

$$\lim_{k\to 0} k^2 R^{ab} K_a K_b \neq 0,$$

along at least one non-spacelike geodesic terminating at the singularity with the value of the affine parameter $k = 0$ at the singularity. The stronger sense in which a strong curvature singularity is defined is given by the requirement that the above limiting condition must be satisfied along *all* non-spacelike geodesics terminating at the naked singularity in the past (Tipler, 1977b). Such a strength was examined earlier for the self-similar case (Waugh and Lake, 1989) along the radial null geodesic which is a root of $V(X) = 0$ (the Cauchy horizon) and also along two other simple null geodesics (Ori and Piran, 1990) that either come out from or fall into the singularity at $(0, 0)$. Our purpose here is to examine the strength of the singularity in the stronger sense stated above and hence we analyse the behaviour of the scalar $\Psi = R^{ab} K_a K_b$ along *all* non-spacelike geodesics that terminate at the naked singularity in the past. Further, both ingoing and outgoing families will be considered. Thus this is a more general analysis which includes earlier results and the behaviour along both the timelike and null cases are covered. For a perfect fluid, the scalar quantity Ψ has the value

$$\Psi = R^{ab} K_a K_b = k_o(\rho + P)e^{2\nu}(K^t)^2 + Bk_o(\rho - P)/2, \tag{6.88}$$

where ρ and P are matter density and pressure. Substituting the values of K^t, ρ, and P in the above we get for the outgoing geodesics,

$$\Psi = \frac{(p + \eta)k_o e^{2\nu}(X + e^{2\psi}Q)^2(C + Bk)^2}{r^4(e^{2\psi} - X^2 e^{2\nu})^2} + \frac{k_o B(\eta - p)}{2r^2}. \tag{6.89}$$

Now, multiply the above by k^2 and take the limit as $k \to 0$. Using l'Hospital's rule and eqns (6.63) and (6.64) and the fact that at the singularity $t = 0, r = 0, k = 0$, and $X = X_o$, we get for the case $V(X_o) = 0$,

$$\lim_{k\to 0} k^2\Psi = \frac{4H_o}{(2+H_o)^2} > 0. \tag{6.90}$$

In a similar manner, it turns out that for the case $Q(X_o) = 0$ we have

$$\lim_{k\to 0} k^2\Psi = H_o U_o^2 > 0, \tag{6.91}$$

where $H_o = H(X_o)$ and $U_o^2 = U^2(X_o)$. It should be noted that along non-spacelike geodesics that terminate at the singularity $H_o > 0$ is satisfied. Along geodesics that terminate with tangent $X = X_o$, which is a root of $Q(X_o) = 0$, a straightforward calculation using eqn (6.65) shows that $U_o^2 > 1$ and hence the strong curvature condition is satisfied along all outgoing geodesics terminating at the singularity.

The results obtained in eqns (6.89) and (6.90) do not change when one considers ingoing non-spacelike geodesics terminating at the singularity. This is done by considering negative sign solutions given by eqns (6.63) and (6.64). Following a similar procedure as above we find that the strong curvature condition is satisfied even along the ingoing geodesics. Clearly this is a strong curvature singularity in a very powerful sense.

Finally, we discuss below in some detail the self-similar Tolman–Bondi dust ($a = 0$) solutions which provide a good illustration of the ideas discussed above. The metric can be written in comoving coordinates as

$$ds^2 = -dT^2 + \frac{r'^2}{1-F}dR^2 + r^2 d\Omega^2,$$

where $r = r(R,T)$, F is a constant with $1 > F \geq 0$, and the dash denotes a differentiation with respect to R. Further, we have

$$\rho = \frac{1}{8\pi r^2 r'} \equiv \frac{\eta(X)}{8\pi R^2},$$

where ρ is the density and $\eta(X)$ is the quantity already used in our earlier discussion.

The marginally bound self-similar Tolman–Bondi solutions ($F = 0$) have been analysed earlier (Eardley and Smarr, 1979; Waugh and Lake, 1988) and provide a good example to illustrate the usefulness of the formalism given here and for comparison with the results already obtained. It is given by

$$r = (9/4)^{1/3} R(\theta - X)^{2/3},$$

where θ is a constant (in the usual notations we have chosen $M(R) = R/2$ and $T_o(R) = \theta R$). To analyse the nature of singularity we write eqn (6.87) for the present case of marginally bound solutions. The quantities involved work out as below:

$$\eta_o = \eta(X_o) = (4/3)(\theta - X_o)^{-1}(3\theta - X_o)^{-1},$$

$$\eta_1 = \frac{\dot{\eta}(X_o)}{\eta(X_o)} = 2(2\theta - X_o)(\theta - X_o)^{-1}(3\theta - X_o)^{-1}.$$

We find $12(\theta - X_o)X_o^3 = (3\theta - X_o)^3$ and hence the fourth order eqn (6.87) turns out to be

$$12X_o^4 - (12\theta + 1)X_o^3 + 9\theta X_o^2 - 27\theta^2 X_o + 27\theta^3 = 0.$$

A quick calculation of the energy parameter H_o shows that

$$H_o = H(X_o) = \frac{16X_o^5}{(\theta - X_o)^8},$$

and hence H_o is always positive for positive X_o. Thus referring to our discussion above, it is immediately concluded that if the singularity is naked (that is, a positive root of eqn (6.72) exists), then it would necessarily be a node where an infinity of integral curves would terminate at the singularity in the past. Further, as shown earlier, this is a strong curvature singularity in a strong sense. The real and positive roots exist for

$$\theta > \frac{13}{3} + \frac{5}{2}\sqrt{3}.$$

This result matches with the earlier work of Eardley and Smarr (1979). An interesting point to note from the above equation is that it admits at least two positive real roots and that for positive θ there are no negative roots. Therefore the singularity is globally naked.

The case $F < 1$ represent bound self-similar solutions and are given in the parametric form by

$$r = \frac{R}{F}(1 - \cos\mu), \quad \mu - X = \frac{1}{F^{3/2}}(\mu - \sin\mu).$$

We get

$$V(X) = \frac{r'^2}{1 - F} - X^2 = 0.$$

To illustrate the situation we choose here $F = 0.01$ and $\theta = 100$. It follows that $V(X)$ has two positive real roots at $X = 98.56, X = 35.6$

The singularity is naked and the families of geodesics meet the strong curvature singularity with tangent $X = 98.56$. It should be noted that the singularity is globally naked as two real positive roots exist. Further, H_o will be positive in this case as well and hence this is also a strong curvature singularity.

6.7 Naked singularity in non-self-similar collapse

The gravitational collapse scenarios considered in the discussion so far, namely the Vaidya–Papapetrou radiation collapse, the perfect fluid incorporating a non-zero pressure, and certain Tolman–Bondi solutions discussed in the last section are all within the framework of a self-similar space-time. Our conclusions on the occurrence and the physical nature of the naked singularity, as indicated by the its curvature strength, are valid for the entire class of general self-similar space-times. Further, as pointed out in the previous section, the relevance of self-similarity is not ruled out for modelling gravitational collapse. However, self-similarity is a strong geometric condition on the space-time and this gives rise to the possibility that the strong curvature naked singularity found there could be a result of this geometric symmetry rather than the gravitational dynamics of the matter therein. In fact, for their considerations on particles creation by shell-focusing singularities, Hiscock, Williams and Eardley (1982) described self-similarity as the most important assumption which they used and expected their results to be substantially different for non-self-similar gravitational collapse. It is thus a matter of importance to learn if occurrence of strong curvature naked singularities are necessarily confined to self-similar space-times only or there are serious examples of naked singularities forming in non-self-similar gravitational collapse as well.

It is the purpose of this section to examine this question. Let us again consider the gravitational collapse of radiation shells described in Section 6.4 by the metric (6.12) and the stress-energy tensor (6.13). The attention there was focused to a linear form of the mass function $m(v)$, which allowed a rather complete specification of details of non-spacelike geodesics families joining the naked singularity and evaluation of the curvature growth along the same. Due to this assumed linearity of $m(v)$, the space-time there is self-similar. However, if $m(v)$ has any other non-linear form, the basic scale invariance required by the self-similarity given by conditions such as $g_{tt}(ct, cr) = g_{tt}(t, r)$ and $g_{rr}(ct, cr) = g_{rr}(t, r)$ is broken and the space-time is no longer self-similar. We shall concern ourselves here with this class of non-self-similar space-times with a general form of mass function. To work out the tangents K^i to non-spacelike geodesics in this case, the components K^θ and K^ϕ are still the same as given by eqns (6.18) and

(6.19), which is a direct consequence of the spherical symmetry of the space-time. The equations for dK^v/dk and dK^r/dk are given by the other two Lagrange equations (where the Lagrangian is $L = \frac{1}{2}g_{ab}\dot{x}^a\dot{x}^b$, the dot being a derivative with respect to the affine parameter k),

$$\frac{dK^v}{dk} + \frac{m(v)(K^v)^2}{r^2} - \frac{\ell^2}{r^3} = 0,$$

$$\frac{dK^r}{dk} + \frac{1}{r}\frac{dm}{dv}(K^v)^2 - \frac{\ell^2}{r^3}\left(1 - \frac{3m(v)}{r}\right) - \frac{Bm(v)}{r^2} = 0.$$

Here B is a constant defined by eqn (6.15). Write now $K^v = P(v,r)/r$, then $K_iK^i = B$ implies

$$k^v = \frac{dv}{dk} = \frac{P}{r}, \tag{6.92}$$

$$k^r = \frac{dr}{dk} = \frac{1 - \frac{2m(v)}{r}}{2r}P - \frac{l^2}{2rP} + \frac{Br}{2P}. \tag{6.93}$$

It is seen using the above equations that the function P satisfies

$$\frac{dP}{dk} = \frac{P^2}{2r^2}\left(1 - \frac{4m(v)}{r}\right) + \frac{l^2}{2r^2} + \frac{B}{2}, \tag{6.94}$$

with the meaning of B as earlier.

While we gave a complete integration of the above equations in the case of $m(v)$ having a linear form in Section 6.4, to specify all the families of non-spacelike geodesics meeting the naked singularity, such details will not be required here. Actually, the details and specification of such families for the case of a general mass function will depend on the specific form of $m(v)$ chosen and subsequent integrations. Here we will first show that even when $m(v)$ has a general non-linear form and we have a non-self-similar collapse situation, a naked singularity forms in the space-time. Then we examine the strength of this singularity along *all* the possible non-spacelike geodesics families which may terminate at the naked singularity in the past to show that this is indeed a strong curvature naked singularity in a powerful sense as in the earlier linear case. Naked singularities in the context of non-self-similar radiation collapse were studied by Lake (1991), who studied the critical direction associated with the node at the origin by constructing a specific mass function. It was shown that the curvature diverges along this null geodesic, which is the Cauchy horizon meeting the naked singularity in the past. This problem was further studied by Joshi and Dwivedi (1992a, 1992b), where the strength of the naked singularity and the curvature growth were examined along non-spacelike curves in general and along *all* the families of non-spacelike geodesics which might terminate at the naked singularity in the past. Also the class of mass function chosen was quite generic. We discuss this approach below. Non-self-similar gravitational collapse of dust giving rise to naked singularity will be studied in the next section.

The radial null geodesics ($\ell = 0, B = 0$) of the space-time are given by eqns (6.92) and (6.93),

$$\frac{dv}{dr} = \frac{2r}{r - 2m(v)}. \tag{6.95}$$

In the case where $m(v) \neq 0$ at $v = 0$, the singularity is surrounded by the horizon and is not naked. This situation corresponds to an initial mass already present at $v = 0, r = 0$, that is, shell collapse in a Schwarzschild background. On the other hand, the situation here is that of radiation injected into an initially flat and empty region and focused into a central singularity of growing mass by a distant spherical source. The source is turned off at a finite time T when the field settles to a Schwarzschild case. It follows that the differential eqn (6.95) has a singular point at $r = 0, v = 0$. The nature of this singular point can be analysed by standard techniques (see for example, Tricomi, 1961; Perko, 1991). Writing

$$2\left(\frac{dm(v)}{dv}\right)_{v=0} = \lambda, \tag{6.96}$$

it can be seen that the singularity is a node provided $0 < \lambda \leq 1/8$. It follows therefore that for this range of values of λ the families of radial null curves meet the singularity with a definite value of the tangent. Using the notation $X = v/r$, two possible values for the same are obtained from eqn (6.95) and l'Hospital rule,

$$X_0 = \left(\frac{v}{r}\right)_{v\to 0, r\to 0} = \left(\frac{dv}{dr}\right)_{v\to 0, r\to 0} = \frac{2}{1 - \lambda X_0}, \tag{6.97}$$

which implies that

$$X_0 = a_{\pm} = \frac{1 \pm \sqrt{1 - 8\lambda}}{2\lambda}. \tag{6.98}$$

It may be noted that when $\lambda = 0$, the structure of the singularity is somewhat complicated and it is not a pure node. It could be a col-node and some characteristics still pass through the singularity which will be naked. This will be discussed in some more detail later.

This establishes the existence of a naked singularity for a class of non-self-similar space-times describing radiation collapse. It is of course possible that the singularity at $v = 0, r = 0$ is only locally naked rather than being globally naked, that is, the strong censorship is violated but not the weak cosmic censorship. It is argued, however, that locally naked singularities should be treated no less seriously as compared to the globally naked singularities (see for example, Penrose, 1979). The reason is that, general relativity being a scale-independent theory, if locally naked singularities can occur on a very small scale, they can also occur on a very large scale when the observer receiving messages from the naked singularity will have enough time to analyse the physics of the same.

In any case, it would be of interest to know whether the naked singularity described here could be globally naked, and if so to specify the conditions for the same. This can be described by an analysis of the critical direction associated with the node, which is the Cauchy horizon given by one of the null geodesics in eqn (6.95). For example, if $r(T) < 2m(T)$ along the Cauchy horizon, the node is only locally naked. AT $v = T$ the collapse stops and the field settles to a Schwarzschild field. So, if the geodesics cross the $v = T$ line with $r(T) < 2m(T)$, they will be ingoing as dv/dr is negative, and geodesics with $r(T) > 2m(T)$ at this line would escape to infinity.

In fact, because of the generality of the mass function involved here, one could always choose T and $M(T)$ such that either no geodesics would cross the $v = T$ line with $r(T) > 2M(T)$ making the singularity only locally naked or otherwise. However, to decide this for a given set of values T and $m(T)$ already chosen, define $\lambda_1 X = 2M(v)/r$. Then eqn (6.95) is written as

$$r = r(X) = C \exp\left(\int \frac{1 - X\lambda_1}{2 - X + \lambda_1 X^2} dX\right), \tag{6.99}$$

where C is a constant and $\lambda_1 = \lambda_1(X)$ is such that at $v = 0$, $\lambda_1 = \lambda$. We can write the above equation in the form

$$r = r(X) = C \exp\left[\int \left(\frac{a'_-}{a'_+ - a'_-}\left(\frac{1}{X - a'_+}\right) - \frac{a'_+}{a'_+ - a'_-}\left(\frac{1}{X - a'_-}\right)\right) dX\right], \tag{6.100}$$

where

$$a'_\pm = \frac{1 \pm \sqrt{1 - 8\lambda_1}}{2\lambda_1}$$

Note that $a'_\pm = a'_\pm(X)$ are functions of X here and reduce to constants when $M(v)$ is linear. Since λ_1 is taken positive and less than or equal to

1/8, we have $a'_{\pm} = a_{\pm}$ and $\lambda_1 = \lambda$ at the singularity. It follows from eqn (6.100) that the geodesic families can meet the singularity with tangent $X = a_+$ and $r = \infty$ can be realized in future along the same geodesic only at $a'_- = X$, that is, the function a'_- approaches this limiting value. This effectively gives the condition for the singularity to be globally naked and would depend on the particular choice of the mass function.

In order to assess the seriousness of this naked singularity, we need to examine the strength of the same as pointed out earlier. We examine here the curvature growth along all possible future directed non-spacelike curves meeting the naked singularity in the past. For a general mass function $m(v)$, the tangents to non-spacelike curves are given by eqns (6.92) and (6.93) where

$$\ell^2 = r^4[(K^\theta)^2 + \sin^2\theta(K^\phi)^2]. \tag{6.101}$$

Again, $K^i K_i = B$ and ℓ and P are arbitrary functions of coordinates. For radial curves we have $\ell = 0$. When the non-spacelike curves are geodesics, ℓ is a constant along a given curve and is identified as the impact parameter and the function P satisfies eqn (6.94). Since, for any family of non-spacelike geodesics meeting the singularity, the tangent has a definite value, dv/dr is well-defined at $v = 0, r = 0$. Then the quantity X_0 given by eqn (6.98) has the value

$$X_0 = \frac{2P_0^2}{P_0^2(1-\lambda X_0) - l_0^2}, \tag{6.102}$$

where

$$P_0 = \lim_{v\to 0, r\to 0} P, \quad l_0 = \lim_{v\to 0, r\to 0} l.$$

To evaluate the scalar $\psi = R_{ij}K^i K^j$, we get, using eqns (6.13) and (6.92) (the parameter $k = 0$ at the singularity and is affine along geodesics)

$$k^2\psi = \frac{2dm(v)}{dv} P^2 \left(\frac{k}{r^2}\right)^2. \tag{6.103}$$

Using eqns (6.92), (6.93) and the l'Hospital's rule, the limit can be evaluated along non-spacelike curves as $k \to 0$ and we get

$$\lim_{k\to 0} k^2\psi = \frac{\lambda X_0^2}{4} \quad \text{for} \quad P_0 \neq \infty, \tag{6.104}$$

$$\lim_{k\to 0} k^2\psi = \frac{\lambda}{(1-\lambda X_0 - H_0)^2} \quad \text{for} \quad P_0 = \infty, \tag{6.105}$$

where

$$H_0 = \lim_{v \to 0, r \to 0} \left(\frac{l^2 + r^2(dP/dk)}{P^2} \right).$$

For the collapse scenario under consideration $g_{ij}T^{ij} = 0$, hence

$$\psi = R_{ij}K^iK^j = T_{ij}K^iK^j.$$

Along non-spacelike curves for which P_0 is finite, or else H_0 is finite, the limiting value of $k^2\psi$ is definite and ψ diverges as $1/k^2$ in the limit of approach to naked singularity. However, the quantity $T_{ij}K^iK^j$ is precisely the local energy density as measured by any observer with a timelike tangent K^a. Hence, any such timelike observer deduces that the energy density blows up very powerfully near singularity, implying this is a strong curvature singularity in an important sense.

When we confine to non-spacelike geodesics families, the strength is measured in the well-defined way described in Section 5.5. For non-spacelike geodesics, eqn (6.94) gives

$$H_0 = (1 - 2\lambda X_0)/2,$$

which is a finite quantity. It thus follows that for any positive value of λ, $k^2 R_{ij}K^iK^j > 0$ is always true near the singularity except for the case $X_0 = 0$. However, this last situation corresponds to $P_0 = 0$ at the singularity. It is not difficult to see by integrating the geodesic equations near the singularity that if a non-spacelike geodesic is meeting the same in the past, this value is not realized along it. This result on strength of singularity generally applies to all the families of non-spacelike geodesics meeting the singularity with a definite tangent, where the singularity could be either locally or globally naked. It follows that the resulting naked singularity is very powerfully strong.

We now discuss the case $\lambda = 0$ in some detail, that is, the derivative $dm(v)/dv$ vanishes at the origin. Then, as seen above, the curvature condition (6.104) or (6.105) is not satisfied. This characterization of the strength implies that in the limit of approach to the singularity, all volume forms along a non-spacelike geodesic are crushed to zero size. However, there are other useful ways in which the strength of a singularity can be tested. One such important criterion is to check whether it is a scalar polynomial singularity (see for example, Hawking and Ellis, 1973). In the following, we show by means of an explicit example that when $\lambda = 0$, even though $R_{ij}K^iK^j$ does not diverge sufficiently fast, the Kretschmann scalar $K = R^{ijkl}R_{ijkl}$ can diverge along a non-spacelike trajectory meeting the singularity in past in the limit of approach to the singularity. Hence, a naked scalar polynomial singularity would result. It is not difficult to see that this represents certain general features of the situation when $m(v)$ is initially non-linear and we analyse the structure of the same in some detail.

To illustrate this clearly, we construct a non-linear mass function which is representative of the class $m(v) \sim v^n$, $n > 1$. This choice here is directed by the requirement that the equation of outgoing non-spacelike curve from the singularity is simple enough which helps towards an easier evaluation of the Kretschmann scalar K near the singularity. Consider $m(v)$ defined by

$$2m(v) = \beta v^{\alpha}\left(1 - 2\alpha\beta v^{\alpha-1}\right), \tag{6.106}$$

where $\alpha > 1$ and $\beta > 0$ is a constant. At $v = 0$, $m(v) = 0$, which gives flat space-time. The null radiation starts imploding at $v = 0$ until $v = T$ where, T satisfies the condition

$$T^{\alpha-1} < \frac{1}{2\beta(2\alpha - 1)}. \tag{6.107}$$

This ensures the positivity of $dm(v)/dv$ and also that $m(v) > 0$. Thus the weak energy condition is satisfied. At $v = T$ we get the Schwarzschild configuration with mass $m_0 = m(T)$.

The radial null geodesics are again given by eqn (6.95), which has a singular point at $v = 0$, $r = 0$. It is seen that for the mass function (6.106), an outgoing radial null geodesic meeting the singularity in past is given by

$$r = \beta v^{\alpha}. \tag{6.108}$$

This non-spacelike curve meets the singularity with a tangent $r = 0$ and it is seen that the singularity is naked. Since $dr/dv > 0$, with increasing v this null geodesic escapes to infinity and the singularity is globally naked. From eqn (6.107) it also follows that the condition $r(T) > 2m(T)$ is satisfied along this trajectory.

One can study the nature of this singularity in general to see that one of the roots of the characteristic equation for eqn (6.95) vanishes because of the fact that

$$\frac{dm(v)}{dv}_{|v=0} = 0.$$

The structure of singularity therefore turns out to be more complicated than the case when $m(v)$ is linear. Further information on the same can be obtained by writing eqn (6.95) in the form (choosing $m(v) \sim v^{\alpha}, \alpha > 1$)

$$U^{\alpha}\frac{dR}{dU} = AR - \frac{U^{\alpha-1}}{2} + 0(R, U). \tag{6.109}$$

Here

$$U = v - 2r, \quad UR = r,$$

A is a positive constant and $0(R, U)$ contains terms of order higher than one. It is seen that (see for example, Tricomi, 1961) the behaviour of integral curves depends on the nature of α. When α is even, the singularity exhibits a col-node structure, that is, given a neighbourhood of the singularity, it behaves like a col for integral curves in a certain region of (v, r) plane and like a node for rest of it. Hence, families of outgoing radial null geodesics can terminate at the singularity in the past in such a case. On the other hand the singularity is a complete node when α is odd. In either case, there are families of integral curves that terminate at the singularity with either $r = 0$ or $v = 2r$ as tangent at the singularity.

Coming to the question of the strength of the singularity, it can be seen that the Kretschmann scalar diverges along all non-spacelike geodesics that meet the naked singularity with a definite tangent. For the case of the mass function given by eqn (6.106), the behaviour of K along the singular curve (6.108) is given by

$$K = 48A \left(\frac{1 - 2\alpha\lambda v^{\alpha-1}}{\lambda^2} \right)^2 \left(\log \frac{Bk}{2\lambda} \right)^{\frac{4\alpha}{\alpha-1}}, \tag{6.110}$$

where A, B are positive constants. Clearly K diverges near the naked singularity as $k \to 0$. In a similar manner it can be shown using the geodesic equations that for any $m(v) \sim v^\alpha$, $\alpha > 1$, the scalar K diverges along all singular non-spacelike geodesics. It is thus seen that even though the singularity is not strong in the sense of the strong curvature condition, it is a strong curvature scalar polynomial singularity.

To conclude, we have shown that a general class of non-self-similar space-times, namely the Vaidya solutions with a non-linear mass term describing radiation collapse, contain a naked singularity which exhibits a strong curvature behaviour. It follows that serious examples of naked singularities are not confined to self-similar space-times only.

6.8 The Tolman–Bondi models

How is the picture given in Section 6.1 for the homogeneous dust ball collapse and the formation of an event horizon altered when inhomogeneities of the matter distribution are taken into account? It is important to include the effects of inhomogeneities because we believe that generally gravitational collapse would start from a very inhomogeneous initial data with a centrally peaked density distribution. Such a study can be made using the Tolman–Bondi class of solutions which model gravitational collapse of an inhomogeneous spherically symmetric dust cloud, that is, a perfect fluid with equation of state $p = 0$ (Tolman, 1934; Bondi, 1948). This is an infinite dimensional family of asymptotically flat solutions of Einstein's equations, which is matched by the Schwarzschild space-time outside the boundary of the collapsing star. The Oppenheimer and Snyder (1939) homogeneous dust ball collapse is a special case of these general class of models.

It can be seen that the introduction of inhomogeneities leads to a rather different picture of gravitational collapse. In fact, as pointed out in Section 6.3, singularities occur in Tolman–Bondi models when shells of dust cross one another at a finite radius, which could be locally and even globally naked. However, such shell-crossing singularities are not regarded as serious counter-examples to censorship hypothesis as discussed earlier. More serious are the shell-focusing singularities occurring on the central world line. Such naked singularities were shown to occur in the case of marginally bound Tolman–Bondi collapse in a numerical study by Eardley and Smarr (1979). Later, Christodoulou (1984) analytically studied the time symmetric case in detail to show the occurrence of shell-focusing naked singularity for the inhomogeneous dust collapse. To assess the seriousness of these examples, Newman (1986a) studied the strength of these naked singularities to show them to be gravitationally weak for both the marginally bound and time symmetric cases, in the sense that the strong curvature condition discussed in the earlier sections is not satisfied along radial null geodesics terminating at the naked singularity in past. It was pointed out, however, by Waugh and Lake (1988), that the class of Tolman–Bondi models considered by Newman and Christodoulou excluded self-similar models. In particular, they considered marginally bound self-similar Tolman–Bondi space-times to show that the strong curvature condition is satisfied along the Cauchy horizon, which is a null geodesic terminating in the naked singularity in the past. Subsequently, Ori and Piran (1990) showed that there are at least two other outgoing null geodesics from the naked singularity along which the strong curvature condition is satisfied. In fact, it follows from the general analysis on self-similar gravitational collapse given in Section 6.5 that for the entire spectrum of self-similar Tolman–Bondi models (the energy content of these models satisfy the energy conditions of Section 5.2), marginally bound or otherwise, there is a non-zero measure set of future directed non-spacelike geodesics which terminate in the naked singularity in the past and the strong curvature condition is satisfied along all these trajectories. It is thus seen that the self-similar Tolman–Bondi collapse provides a serious counter-example to cosmic censorship as far as the curvature strength criterion is concerned.

We now further analyse here the structure of naked singularity forming in an inhomogeneous dust cloud collapse to show that the occurrence of strong curvature naked singularity is not confined to self-similar space-times or radiation collapse only. This is done by pointing out a wide class of Tolman–Bondi models which is non-self-similar in general where such a singularity forms. This also generalizes the earlier work in the sense that all the self-similar Tolman–Bondi space-times containing a strong curvature naked singularity are included here as a special case. The results of Sections 6.4 and 6.5 dealt when the form of the matter was either an inflowing radiation or a perfect fluid with non-zero pressure, however, with the geometric restriction of self-similarity. This self-similarity restriction was relaxed in Section 6.6 for the case of matter being the inflowing radiation. Now it is shown here that these results can be generalized even to the case when the matter is no longer in the form of radiation but is a pressureless dust, and the space-time need not be self-similar.

We use the comoving coordinates (t, r, θ, ϕ) to describe the spherically symmetric collapse of an inhomogeneous dust cloud. The coordinate r has non-negative values and labels the spherical shells of dust and t is the proper time along the world lines of dust particles given by $r =$ const. The collapse of spherically symmetric inhomogeneous dust is described by the Tolman metric in co-moving coordinates (that is, $u^i = \delta^i_t$) and is given by

$$ds^2 = -dt^2 + \frac{R'^2}{1+f}dr^2 + R^2(d\theta^2 + sin^2\theta\, d\phi^2), \tag{6.111}$$

$$T^{ij} = \epsilon\delta^i_t\delta^j_t, \quad \epsilon = \epsilon(t,r) = \frac{F'}{R^2R'}, \tag{6.112}$$

where T^{ij} is the stress-energy tensor, ϵ is the energy density, and R is a function of both t and r given by

$$\dot{R}^2 = \frac{F}{R} + f. \tag{6.113}$$

Here the dot and the prime denote partial derivatives with respect to the parameters t and r respectively and as we are only concerned with the gravitational collapse of dust, we require

$$\dot{R}(t,r) < 0.$$

The quantities F and f are arbitrary functions of r. The quantity $4\pi R^2(t,r)$ gives the proper area of the mass shells and the area of such a shell at $r =$ const. goes to zero when $R(t,r) = 0$. Integration of eqn (6.113) gives

$$t - t_0(r) = -\frac{R^{3/2}G(-fR/F)}{\sqrt{F}}, \tag{6.114}$$

where $G(y)$ is a strictly real positive and bound function which has the range $1 \geq y \geq -\infty$ and is given by

$$
\begin{aligned}
G(y) &= \left(\frac{\sin^{-1}\sqrt{y}}{y^{3/2}} - \frac{\sqrt{1-y}}{y}\right) \quad \text{for} \quad 1 \geq y > 0, \\
G(y) &= \tfrac{2}{3} \quad \text{for} \quad y = 0, \\
G(y) &= \left(\frac{-\sinh^{-1}\sqrt{-y}}{(-y)^{3/2}} - \frac{\sqrt{1-y}}{y}\right) \quad \text{for} \quad 0 > y \geq -\infty,
\end{aligned}
\tag{6.115}
$$

and $t_0(r)$ is a constant of integration. We thus have in all three arbitrary functions of r, namely $f(r)$, $F(r)$, and $t_0(r)$. One could, however, use the remaining coordinate freedom left in the choice of scaling of r in order to reduce the number of such arbitrary functions to two. We therefore rescale R using this coordinate freedom such that

$$R(0, r) = r. \tag{6.116}$$

Then $t_0(r)$ is evaluated by using the equation above and eqn (6.114) to give

$$t_0(r) = \frac{r^{3/2} G(-fr/F)}{\sqrt{F}}. \tag{6.117}$$

The time $t = t_0(r)$ corresponds to $R = 0$ where the area of the shell of matter at a constant value of the coordinate r vanishes. It follows that the singularity curve $t = t_0(r)$ corresponds to the time when the matter shells meet the physical singularity. Thus, the range of the coordinates is given by

$$0 \leq r < \infty, \qquad -\infty < t < t_0(r). \tag{6.118}$$

It follows that unlike the collapsing Friedmann case, where the physical singularity occurs at a constant epoch of time (say, at $t = 0$), the singular epoch now is a function of r as a result of inhomogeneity in the matter distribution. One could recover the Friedmann case from the above equations if we set

$$t_0(r) = t_0'(r) = 0.$$

The function $f(r)$ classifies the space-time as bound, marginally bound, or unbound depending on the range of its values which are

$$f(r) < 0, \quad f(r) = 0, \quad f(r) > 0,$$

respectively. The function $F(r)$ can be interpreted as the weighted mass (weighted by the factor $\sqrt{1+f}$) within the dust ball $\mathcal{B}$ of coordinate radius r which is conserved in the following sense.

$$m(r) = \frac{F(r)}{2} = \int_{\mathcal{B}} (1+f)^{1/2} \epsilon(t,r) dv = 4\pi \int_0^r \rho(r) r^2 dr, \qquad (6.119)$$

where $\epsilon(0,r) = \rho(r)$. For physical reasonableness the weak energy condition would be assumed throughout the space-time, that is, $T_{ij}V^iV^j \geq 0$ for all non-spacelike vectors V^i. This implies that the energy density ϵ is everywhere positive, ($\epsilon \geq 0$) including the region near $r = 0$.

Using the scaling given by eqn (6.116), the energy density ϵ on the hypersurface $t = 0$ is written as

$$\epsilon = \frac{F'}{r^2}.$$

Since the weak energy conditions are satisfied and F is a function of r only, it follows that $F' \geq 0$ throughout the space-time.

Singularities are the points of space-time where the normal differentiability and manifold structures break down. In other words, points where the energy density given by eqn (6.112), or the curvature quantities such as the scalar polynomials constructed out of the metric tensor and the Riemann tensor diverge. One example of such a quantity is the Kretschmann scalar $\mathcal{K} = R_{abcd}R^{abcd}$, which is given in the Tolman–Bondi case by

$$\mathcal{K} = 12\frac{F'^2}{R^4R'^2} - 32\frac{FF'}{R^5R'} + 48\frac{F^2}{R^6}. \qquad (6.120)$$

Such singularities are indicated by the existence of incomplete future or past directed non-spacelike geodesics in the space-time which terminate at the singularity. Then one requires that the curvature quantities stated above assume unboundedly large values in the limit of approach to the singularity along the non-spacelike geodesics terminating there. If such a condition is satisfied, then one would like to consider the singularity to be a physically significant curvature singularity.

In Tolman–Bondi space-times singularities occur, as one can see from eqns (6.112) and (6.120), at points where $R = 0$, which are called shell-focusing singularities, and also at points where $R' = 0$. At the points where $R' = 0$ the Tolman–Bondi metric is degenerate and these are called shell-crossings. In the context of Tolman–Bondi space-times the points $R > 0, F' > 0$, where $R' = 0$, are called the shell-crossing singularities (Newman, 1986a). In the present consideration, we do not consider such shell crossings and assume that there are no shell-crossing singularities in the space-time, except probably right at the center $r = 0$ (Eardley, 1987). Actually, the Tolman–Bondi models are fully characterized in terms of the behaviour of the functions f and F and hence the conditions for the absence of shell crossing singularities for $r > 0$ can be given in terms of the behaviour of these functions in the range $r_c \geq r > 0$, r_c being the boundary of the dust cloud (see for example, Hellaby and Lake, 1985). Such shell-crossing singularities in Tolman–Bondi space-times have been analysed in detail in the literature (Muller zum Hagen, Yodzis and Seifert, 1973, 1974), and their nature appears to be fairly well understood. Even though they could be locally naked, the important point is they have been shown to be gravitationally weak (Newman, 1986a). Thus it is generally believed that such shell-crossing singularities need not be taken seriously as far as the cosmic censorship conjecture is concerned. The absence of shell-crossing singularities in a space-time turns out to be related to the condition that the function $t_0(r)$ giving the proper time for the shells to fall into the physical singularity should be a monotonically increasing function. The dust density and certain components of the curvature blow up near such a singularity. However, the causal structure of the space-time can be extended through such a singularity and the space-time metric also can be defined in the neighbourhood of such a point in a distributional sense (Papapetrou and Hamoui, 1967). We note, however, that such an extension need not be unique or even be dust.

Unlike the shell-crossings, the space-time metric admits no extension through a shell-focusing singularity occurring at $R = 0$, which is more difficult to ignore. A shell-focusing singularity can be avoided only by rejecting the forms of matter such as dust as the fundamental forms of matter. Hence, we concentrate our attention here only on such shell-focusing singularities and examine their nature and structure for the Tolman–Bondi space-times. It has been shown earlier (Christodoulou, 1984) that a shell-focusing singularity occurring at $r > 0, R = 0$ is totally spacelike and therefore our discussion would be confined to the singularity at $r = 0$.

The points (t_0, r_0) where a shell-focusing singularity $R(t_0, r_0) = 0$ occurs, are related by eqn (6.114). Here the singularity is said to occur at $r = r_0$ at coordinate time $t = t_0$ and we would call the singularity to be a *central singularity* if it occurs at $r = 0$. Earlier work (Christodoulou, 1984; Newman, 1986a) has shown that this central shell-focusing singularity is naked, though gravitationally weak, for a class of Tolman–Bondi space-time for which the energy density (which is assumed to be positive everywhere and is taken to be non-zero at $r = 0$) and the metric are even smooth functions of t and r. In terms of functions $F(r)$ and $f(r)$ it amounts to the conditions

$$F(r) = r^3\mathcal{F}(r), \qquad \infty > \mathcal{F}(0) > 0, \qquad 0 < p(r) \leq 1. \tag{6.121}$$

It was, however, pointed out by Waugh and Lake (1988) and Ori and Piran (1990) that this class of gravitationally weak naked singularities excludes the self-similar Tolman–Bondi models, where they showed the singularity to be gravitationally strong along the Cauchy horizon, which is a null geodesic coming out of the singularity. Further, Grillo (1991) pointed out an example in the case of unbound Tolman–Bondi models which are non-self-similar and the naked singularity is gravitationally strong. In fact, Dwivedi and Joshi (1992) have pointed out that the naked singularity exists and is gravitationally strong for a wide class of Tolman–Bondi models which are non-self-similar in general and include all the self-similar models as a special subclass.

In order to represent the gravitational collapse situation, we assume the energy density ϵ to have compact support on an initial spacelike hypersurface and the Tolman–Bondi space-times given by eqn (6.111) can be matched at some $r = \text{const.} = r_c$ to the exterior Schwarzschild field

$$ds^2 = -\left(1 - \frac{2M}{r_s}\right) dT^2 + \frac{dr_s^2}{1 - \frac{2M}{r_s}} + r_s^2 d\Omega^2, \tag{6.122}$$

where $d\Omega^2 = d\theta^2 + sin^2\theta \, d\phi^2$. The value of the Schwarzschild radial coordinate is $r_s = R(t, r_c)$ at the boundary $r = r_c$. We have $m(r_c) = M$, where M is the total Schwarzschild mass enclosed within the dust ball of coordinate radius of $r = r_c$. Without going into further details of the matching conditions we would like to say a few words regarding the apparent horizon. The apparent horizon in the interior dust ball lies at $R = F(r)$. From eqns (6.114) and (6.117) one can see that the corresponding time $t = t_{ah}(r)$ is given by

$$t = t_{ah}(r) = \frac{r^{3/2} G(-p)}{\sqrt{F}} - FG(-f). \tag{6.123}$$

It is seen that emissions from the shell-focusing singularity $R(t_0, r_0) = 0$ for all $r_0 > 0$ would lie in the region above $t = t_{ah}$ that is, $t_0 > t_{ah}$ for all $r_0 > 0$, t_0 being the time when singularity at $r = r_0$ occurs. Hence all radiations would be future trapped from shell-focusing singularities at $r > 0$. At $r = 0$, however, $t(0) = t_{ah}(0)$ and the singularity could be at least locally naked. Any light ray terminating at this singularity in the past goes to the future infinity if it reaches the surface of the cloud $r = r_c$ earlier than the apparent horizon at $r = r_c$. In such a case the singularity would be globally naked. We now examine this central singularity below.

Partial derivatives of R like R' and $\dot{R}'$ are of crucial importance in our analysis. Hence for all the cases of collapse, namely the bound, marginally bound and unbound cases ($f < 0, f = 0$ and $f > 0$) it is convenient to have their expressions, which are given using the eqns (6.113) to (6.117) as

$$\left(\frac{\partial R(t,r)}{\partial r}\right)_{t=\text{const.}} = R' = (\eta-\beta)P - \left[\frac{1+\beta-\eta}{\sqrt{\lambda+f}} + (\eta - \tfrac{3}{2}\beta)\tfrac{t}{r}\right]\dot{R}, \tag{6.124}$$

$$\left(\frac{\partial R'(t,r)}{\partial t}\right)_{r=\text{const.}} = \frac{\beta}{2r}\dot{R} + \frac{\lambda}{2rP^2}\left[\frac{1+\beta-\eta}{\sqrt{\lambda+f}} + (\eta - \tfrac{3}{2}\beta)\tfrac{t}{r}\right], \tag{6.125}$$

$$t_0' = \frac{1+\beta-\eta}{\dot{R}(0,r)} + \frac{t_0}{r}(\eta - \tfrac{3}{2}\beta), \tag{6.126}$$

where we have put

$$R(t,r) = rP(t,r), \quad \eta = \eta(r) = r\frac{F'}{F}, \quad \beta = \beta(r) = r\frac{f'}{f}, \quad F(r) = r\lambda(r). \tag{6.127}$$

At this point we would restrict ourselves to functions $f(r)$ and $\lambda(r)$ which are analytic at $r=0$ such that

$$\lambda(0) \neq 0. \tag{6.128}$$

It should be pointed out that in case $\lambda(r) = \text{const.}$ and $f(r) = \text{const.}$ the space-time becomes self-similar in the sense of

$$g_{tt}(t,r) = g_{tt}(ct,cr), \quad g_{rr}(t,r) = g_{rr}(ct,cr)$$

and

$$R(ct,cr) = cR(t,r),$$

for any $c > 0$, whereas in general it is non-self-similar.

The condition that $\lambda(0) \neq 0$ implies that $t_0(0) = 0$. It follows that the point $r=0, t=0$ corresponds to the central singularity on the $t=0$ hypersurface where the energy density ϵ blows up. It is useful to note the difference between the considerations here and those in earlier works, such as Newman (1986a). The major difference between these two classes of models is of course clear, which is that while with the scaling of coordinates given by eqn (6.116) the density and other functions are smooth and even on the $t=0$ hypersurface in the case of Newman, in the present class there is a singularity at $r=0$ on this surface where the density blows up. It follows that the class being treated here is different from models treated earlier, and it admits a strong curvature singularity as will be shown below. To see this difference further, one could examine the behaviour of physical quantities such as the density at an epoch of time before the occurrence of central singularity as was done by Ori and Piran (1990) for the case of marginally bound self-similar Tolman–Bondi models. It is seen without difficulty in that case, and also in the present case, that the density function is non-analytic, though non-singular, at $r = 0$ at these earlier epochs. It is this crucial difference between the models considered earlier, which gave rise to a weak singularity there, and the models considered here (and also in Waugh and Lake (1989) and Ori and Piran (1990) for the self-similar Tolman–Bondi models) which gives a strong curvature naked singularity to be seen below. The $t = 0$ hypersurface is singular here at $r = 0$ and hence the non-singular initial data for collapse can be specified, for example in the case of $f = 0$ marginally bound collapse, on any initial slice $t < 0$ for the situation under consideration.

At this point we introduce a new variable

$$X = \frac{t}{r}. \tag{6.129}$$

Then the function $P(t, r) = P(X, r)$ is given with the help of eqns (6.114) and (6.117)

$$X - \Theta = -\frac{P^{3/2}}{\sqrt{\lambda}} G(-fP/\lambda), \tag{6.130}$$

where we have put

$$t_0 = r\Theta(r) = rG(-f/\lambda)/\sqrt{\lambda}.$$

The tangents $K^r = dr/dk$ and $K^t = dt/dk$ to the outgoing radial null geodesics with k as the affine parameter satisfy

$$\frac{dK^t}{dk} + \dot{R}'\sqrt{1+f}K^r K^t = 0, \tag{6.131}$$

$$\frac{dt}{dr} = \frac{K^t}{K^r} = \frac{R'}{\sqrt{1+f}}. \tag{6.132}$$

Our main purpose now is to find whether these geodesics terminate at the central singularity formed at $r = 0, t = t_0(0)$ in the past. The exact nature of this singularity $t = 0, r = 0$ could be analysed by the limiting value of $X = t/r$ at $t = 0, r = 0$. If the geodesics meet the singularity with a definite value of the tangent then using eqn (6.132) and l'Hospital rule we get

$$X_0 = \lim_{t\to 0, r\to 0} \frac{t}{r} = \lim_{t\to 0, r\to 0} \frac{dt}{dr} = \lim_{t=0, r=0} \frac{R'}{\sqrt{1+f}}. \tag{6.133}$$

We use the notation

$$\lambda_0 = \lambda(0), \quad \beta_0 = \beta(0), \quad f_0 = f(0), \tag{6.134}$$

and $Q = Q(X) = P(X, 0)$, which is a function of X alone given implicitly by eqn (6.130)

$$X - \Theta_0 = -\frac{Q^{3/2} G(-f_0 Q/\lambda_0)}{\sqrt{\lambda_0}}. \tag{6.135}$$

Here $\Theta = \Theta(0)$ and we would denote $Q_0 = Q(X_0)$ throughout. We can simplify eqn (6.133) with the help of eqns (6.114) to (6.117), (6.124) to (6.126) and (6.135) as

$$V(X_0) = 0, \tag{6.136}$$

where

$$V(X) \equiv (1 - \beta_0)Q + \left(\frac{\beta_0}{\sqrt{\lambda_0 + f_0}} + (1 - \frac{3}{2}\beta_0)X\right)\sqrt{\frac{\lambda_0}{Q} + f_0} - X\sqrt{1 + f_0}. \tag{6.137}$$

Hence if the equation $V(X) = 0$ has a real positive root, the singularity could be naked. In order to be the end point of null geodesics at least one real positive value of X_0 should satisfy eqn (6.136). Clearly, if no real positive root of the above is found, the singularity $t = 0, r = 0$ is not naked. It should be noted that many real positive roots of the above equation may exist which give the possible values of tangents to the singular null geodesics terminating at the singularity. However, such integral curves may or may not realize a particular value X_0 at the singularity. To determine whether a value X_0 is realized along any outgoing singular geodesics to give a naked singularity, consider the equation of the radial null geodesics in the form $r = r(X)$. From eqns (6.132) we have

$$\frac{dX}{dr} = \frac{1}{r}\left(\frac{dt}{dr} - X\right) = \frac{R' - X\sqrt{1+f}}{r\sqrt{1+f}}. \tag{6.138}$$

Using eqns (6.124) to (6.126) and eqn (6.137) and the fact that functions $\lambda(r)$ and $f(r)$ are analytic, we could write eqn (6.138) as

$$\frac{dX}{dr} = \frac{V(X)}{r\sqrt{1 + f_0}} + \frac{S(X, r)}{r}. \tag{6.139}$$

Here $S(X, r)$ is some function of X and r such that at $r = 0, S(X, 0) = 0$.

Suppose now that X_0 is a simple root of the equation $V(X) = 0$. Therefore with the help of eqn (6.137) we could write

$$V(X) = h_0\sqrt{1 + f_0}(X - X_0) + \text{terms of higher order in}(X - X_0), \tag{6.140}$$

where

$$h_0 = \frac{1}{\sqrt{1+f_0}}\left(\frac{\lambda_0}{2Q_0^2}\left(\frac{\beta_0}{\sqrt{\lambda_0+f_0}} + X_0(1-\frac{3}{2}\beta_0)\right) - \frac{\beta_0}{2}\sqrt{\frac{\lambda_0}{Q_0}+f_0}\right) - 1.$$

Substituting eqn (6.140) in eqn (6.139) we get

$$\frac{dX}{dr} - (X - X_0)\frac{h_0}{r} = \frac{H}{r}. \tag{6.141}$$

where $H = H(X, r)$ contains terms of higher powers of $(X - X_0)$ and positive powers of powers of r and is such that $H(X_0, 0) = 0$, that is, at the point $r = 0, X = X_0, H \to 0$.

Integration of eqn (6.141) gives the equation of geodesics as $r = r(X)$. Multiplying eqn (6.141) by r^{-h_0} and integrating gives

$$X - X_0 = Cr^{h_0} + r^{h_0}\int Hr^{(-h_0-1)}dr, \tag{6.142}$$

where C is a constant of integration that labels different geodesics. If the singularity is the end point of these geodesics with tangent $X = X_0$, we must have $X \to X_0$ as $r \to 0$ in eqn (6.142). Note that as $X \to X_0, r \to 0$, the last term in eqn (6.142) always vanishes near the singularity regardless of the constant h_0 being either positive or negative. Therefore the single null geodesic described by $C = 0$ always terminates at the singularity $t = 0, r = 0$, with $X = X_0$. In the case where $h_o > 0$ there are infinitely many integral curves characterized by different values of C that terminate at the singularity. However, in the case $h_0 < 0$ there is only one singular geodesic characterized by $C = 0$ that terminates at the singularity.

It is thus seen that the existence of a positive real root of eqn (6.136) is a necessary and sufficient condition for the singularity to be at least locally naked. Such a singularity could be globally naked as well. The details of this will depend on the global features of $\lambda(r)$. A globally naked shell-focusing singularity forming in an inhomogeneous dust cloud collapse is shown in Fig. 52.

We point out below a special class of generally non-self-similar dust which illustrates some of these points. Consider the subclass defined by the condition

$$f(r) = \text{const.} = 0, \quad \lambda_0 \neq 0. \tag{6.143}$$

We have, from eqns (6.117) and (6.135)

$$\Theta_0 = \frac{2}{3\sqrt{\lambda}}, \quad Q(X) = \Theta_0^{2/3}(X - \Theta_0)^{2/3}. \tag{6.144}$$

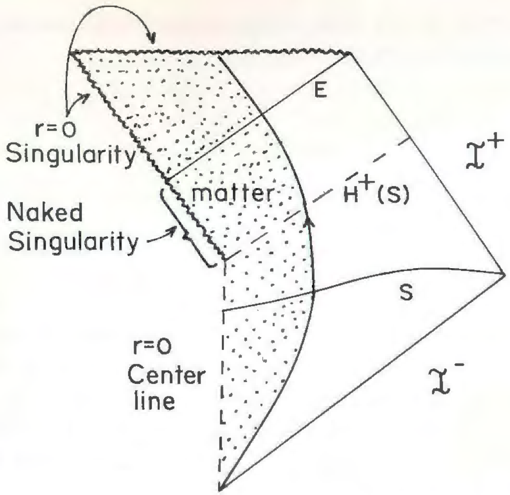

Fig. 52 A shell-focusing naked singularity occurring in the gravitational collapse of inhomogeneous dust. The surface $H^+(S)$ is the Cauchy horizon and E is the event horizon. The dotted region denotes the collapsing dust.

Equations (6.136) and (6.144) imply that

$$(3\Theta_0 - X_0)^3 = 27X_0^3\Theta_0^2(\Theta_0 - X_0). \tag{6.145}$$

The above has real and positive roots if

$$\Theta_0 \geq \frac{2}{9}\left(\frac{25}{3} + 5\sqrt{3}\right), \tag{6.146}$$

and therefore the singularity is naked for space-times for which eqn (6.146) is satisfied. Note that in case $\lambda(r) = \lambda_0$ the space-time becomes self-similar and the result obtained above agrees with that of Eardley and Smarr (1979). Further eqn (6.145) has two real positive roots. Our analysis of self-similar collapse in Section 6.5 shows that in this case the singularity will be globally naked.

Finally, we need to determine the curvature strength of the naked singularity at $t = 0$ and $r = 0$. For this purpose the strong curvature condition discussed earlier must be verified in the limit of approach to the singularity, namely

$$\lim_{k\to 0} k^2 R_{ab}K^aK^b > 0. \tag{6.147}$$

This provides a sufficient condition for all the two-forms $\mu(k)$ defined along the singular null geodesic to vanish in the limit of approach to the singularity and implies a very powerful curvature divergence establishing a strong curvature singularity.

For the Tolman–Bondi space-times under consideration, one could use eqns (6.113) and (6.115) in order to get

$$\Psi \equiv R_{ab}K^a K^b = \frac{F'(K^t)^2}{R^2R'} = \frac{F'(K^t)^2}{r^2P^2R'}. \tag{6.148}$$

Using l'Hospital rule and eqns (6.124) to (6.126), (6.131) and (6.132) we get

$$\lim_{k\to 0} k^2\Psi = \frac{\lambda_0 X_0}{\sqrt{1+f_0(h_0+2)^2Q_0^2}} \neq 0. \tag{6.149}$$

Therefore the strong curvature condition is satisfied. It should be noted that the value $X_0 = 0$ is not realized along the singular geodesics in the limit of approach to the singularity. This is not difficult to see. From the eqn (6.116) and the definition of the the quantity P given by $R = rP$ it follows that at $X = 0$, that is, $t = 0$, we have $P(0, r) = 1$. Now, $Q(X) = P(X, 0)$ implies that for $X = 0$ we have

$$Q(0) = Q_0 = P(0, 0) = 1.$$

Using this fact one sees clearly that $V(0) = 1$, therefore $X = 0$ cannot be the root of the equation $V(X_0) = 0$. This fact can be derived independently from eqn (6.135) as well.

We note finally that the class of models considered here is actually a part of a still wider class of Tolman–Bondi space-times given by the functions $F(r), f(r)$ being at least C^1. These classes could be discussed similarly; however, the discussion will be more complicated in view of the generality of the situation.

A limitation of the models considered here could be that the pressure has been assumed to be vanishing, $p = 0$, which could be important in the final stages of the collapse. It is possible, on the other hand, that in the final stages of collapse, the dust equation of state could be relevant (see for example, Penrose (1974a), Hagerdorn (1968)), and at higher and higher densities the matter may behave more and more like dust. Further, if there are no large negative pressures (as implied by the validity of the energy conditions), then the pressure also might contribute in a positive manner gravitationally to the effect of the dust and may not alter the conclusion derived here, as in the case of the perfect fluid considered in Section 6.5.

We have investigated here the occurrence and nature of naked singularity for the inhomogeneous gravitational collapse of Tolman–Bondi dust clouds. It is shown that the naked singularities form in a very wide class of collapse models which includes the earlier cases considered by Eardley and Smarr (1979) and Christodoulou (1984). This class also contains self-similar as well as non-self-similar models. Subclasses are identified here when the naked singularity will be gravitationally strong as discussed earlier. The structure and strength of this singularity have been examined. The analysis presented here should be useful for any possible rigorous formulation of the cosmic censorship hypothesis. An implication of the results of this as well as the earlier section for the fundamental issue of the final fate of gravitational collapse is that naked singularities need not be considered as artifacts of geometric symmetries of space-time such as self-similarity. It may be that the existence of naked singularity is not just a geometric phenomena and the answer to cosmic censorship conjecture could lie in the dynamics of the Einstein equations. Of course, one could argue that the origin of occurrence of such naked singularities could be basically the use of matter fields such as the dust and perfect fluid. Though used very widely in astrophysical approximations, such matter forms may create singularities even in Newtonian gravity. Hence, such naked singularities may not be an attribute of the general theory of relativity. We will discuss this issue further in the next chapter.

7

GENERAL CONSTRAINTS ON NAKED SINGULARITY FORMATION

In the previous chapter, we showed that strong curvature naked singularities arise in a variety of situations involving dust, perfect fluid, and radiation collapse for self-similar as well as non-self-similar space-times. These are dynamical situations representing a gravitational collapse scenario and the matter content of the space-time is reasonable in that the energy conditions implying the positivity of energy are satisfied and widely used equations of state are obeyed.

Hence, these collapse scenarios must be taken into account for any possible formulation of the cosmic censorship hypothesis. In fact, in order to arrive at any provable formulation of the censorship hypothesis, or otherwise to assess the reasonability of these examples, it appears necessary to examine possible constraints which must be obeyed by naked singularities occurring in the general theory of relativity. It is the purpose of this chapter to derive such constraints on naked singularities and to examine various properties of the same which may give a better insight into their occurrence and nature. In Section 7.1 we discuss the conclusions on the structure of naked singularity that follow from the results available so far. While discussing the spectrum of possibilities and conclusions arising from the analysis of gravitational collapse in Chapter 6, it is also shown here that the conclusions on self-similar collapse of a perfect fluid discussed in Section 6.5 can in fact be generalized for any form of matter satisfying the weak energy condition for the positivity of mass-energy density. The relationship between the causal structure distortions that a naked singularity would cause in the space-time, and its strength, is discussed in Section 7.2 based on investigations carried out on this topic. Topology change in a space-time is examined in Section 7.3 and is shown to give rise to naked singularities. An important open issue concerning the naked singularities is the stability of the same when they form in gravitational collapse. We discuss this problem in Section 7.4. Finally, in Section 7.5, non-spherical collapse is discussed and other conjectures providing an alternative to cosmic censorship, such as the hoop conjecture of Thorne and the event horizon conjecture of Israel are discussed.

7.1 The structure of naked singularity

The singularity theorems discussed in Chapter 5 establish the existence of space-time singularities in the form of incomplete non-spacelike geodesics in the space-time for both the situations of gravitational collapse and cosmology. These theorems, however, provide no information on the nature of such singularities such as whether they occur in the past or future, the possible growth of curvature in the limit of approach to the singularity, and whether they will always be covered by an event horizon hidden from all outside observers. Thus, while dealing with the gravitational collapse scenarios it becomes crucial to assume the cosmic censorship in the form of future asymptotic predictability of the space-time which ensures that the resulting singularities are necessarily hidden behind an event horizon. This assumption is basic to the development and applications of physics of black-holes.

We have been concerned in the previous chapter with the nature of such singularities arising in gravitational collapse of massive bodies such as stars. Such a situation is characterized by the condition that the space-time contains a regular spacelike hypersurface on which regular initial data is defined and the stress-energy tensor has a compact support. All physical quantities are assumed to be well-behaved on such an initial hypersurface. We are concerned with the dynamical evolution of this initial data into the future which determines the final fate of the collapsing configuration. The collapsing matter is supposed to give rise to a singularity in the sense that there are future directed non-spacelike curves from the initial surface which have finite length and are future incomplete in the sense described earlier. We have now examined the question whether there exists future directed non-spacelike trajectories in such a space-time which are past incomplete and terminate at the singularity in the past; that is, whether such a collapse could give rise to naked singularities as well. The existence of such a causal curve establishes the existence of a naked singularity which is at least locally naked. If this trajectory reaches an observer at infinity, the singularity will be globally naked as well. In fact, there are many static space-time situations known in Einstein's gravity which contain naked singularity. These, however, are not relevant to the issue of the final fate of gravitational collapse which is our concern here.

Our examination of such dynamical collapse scenarios, where we are concerned with the evolution of regular initial data from a well-behaved initial value surface imply the following basic conclusions. Firstly, such a naked singularity forms in the dynamical evolution of several forms of matter such as the collapse of inflowing radiation, dust or perfect fluid. Secondly, a non-zero measure set of non-spacelike trajectories, in the form of families of non-spacelike curves, are emitted from the naked singularity as opposed to a single null geodesic escaping which corresponds to a single wave front coming out. Finally, such a singularity is physically significant in the sense that it is a powerfully strong curvature singularity in that the curvatures diverge rapidly along all the trajectories meeting the naked singularity in the past.

How seriously such a naked singularity is to be taken, and what are the implications of the same towards the formulation and proof of the cosmic censorship hypothesis? It may be noted that all the gravitational collapse situations we have investigated so far are spherically symmetric. Is it possible that the naked singularities occurring are the artifacts of this assumed symmetry? Though this possibility cannot be ruled out, it is relevant to note here the numerical results of Shapiro and Teukolsky (1991), showing that naked singularities could occur in the non-spherical prolate models of collapse of collisionless particles as well. Thus their existence need not be due to the assumed spherical symmetry only. Further, as was shown by the singularity theorems, the singularities developing in the spherically symmetric situations still persist even when small perturbations from this symmetry are taken into account. It is possible that a similar situation may arise here and in this context a detailed investigation of the spherically symmetric scenario becomes quite important.

Coming to the specific situations we have investigated, the Vaidya radiation models may appear somewhat artificial in their matter content in that these models involve the collapse of inflowing directed radiation with the equation of state $T_{ij} = \sigma k_i k_j$. The space-time, however, satisfies the weak as well as the strong energy conditions and it is conceivable that in the very final stages of collapse all the matter of the star is relativistic and largely radiation dominated, which is flowing inwards radially in a directed manner due to the very strong gravitational pull. Further, for such an inflowing radiation, the radiation pressure equals the density of the flowing radiation and hence the stress-energy tensor includes both the energy density and pressure components. In fact, as it follows from the analysis of Sections 6.4 and 6.7 and as discussed by Waugh and Lake (1989) and Lemos (1992), the basic features of the collapse and the resulting naked singularity are similar for the Vaidya collapse and the self-similar Tolman–Bondi dust solutions, and in a sense the radiation collapse can be viewed as a limiting situation of the collapsing dust.

The basic advantage, however, with the radiation collapse scenario is that it is possible to fully understand there the structure of the naked singularity forming and the families of non-spacelike curves terminating at the naked singularity in the past in detail, as we have shown in Section 6.4. This allows an important question of principle to be settled, namely that strong curvature naked singularities can develop under certain situations in the gravitational collapse in general relativity, from a non-singular initial data defined in terms of matter satisfying the energy conditions. As was made clear in Section 6.3, there are now many forms of cosmic censorship hypothesis, and it is a matter of importance to determine which might conceivably succumb to a proof. To this end, it is necessary to understand as fully as possible the structure of the available examples of naked singularities, and the analysis of Section 6.4 is directed towards this purpose. The usefulness of this analysis from the point of view of achieving a provable formulation of censorship hypothesis is clear from the fact that the Vaidya–Papapetrou models provide a counter-example to certain forms of censorship hypothesis. We may note that this analysis concerns the outgoing causal curves from the naked singularity, because the ingoing ones lie in a Minkowskian region and so are not subjected to any curvature. Such features are worth noting because they indicate how censorship conjectures may be reformulated to accommodate the radiation collapse scenario and other available examples.

The radiation collapse space-times considered in Section 6.4 are self-similar as the mass function involved is linear. We generalize these results on the occurrence and strength of naked singularity to the general class of self-similar space-times in Section 6.5, where we consider a perfect fluid collapse rather than the inflowing radiation. This general result shows that the self-similarity of a space-time is not compatible with the cosmic censorship hypothesis for a wide range of physically reasonable matter including the inflowing radiation, dust, and the perfect fluid with an adiabatic equation of state. Thus, cosmic censorship rules out the possibility of a self-similar collapse. On the other hand, this also raises the question whether the naked singularities basically arise due to the self-similarity of the space-times. The answer here turns out to be in the negative as our further considerations on gravitational collapse have shown that strong curvature naked singularities arise for several non-self-similar space-times as well. The Vaidya space-times with a non-linear mass function are no longer self-similar and provide a class which could be relevant for the very final stages of collapse. Similarly, the non-self-similar collapse of dust also gives rise to powerful strong curvature singularities as seen by our investigation of the Tolman–Bondi models.

The criterion defined for the strength of naked singularity in Section 5.4 and used here to examine the strength of naked singularities is of course subject to refinement. However, it certainly implies a very powerful curvature divergence near the naked singularity. In fact, for the case of the radiation collapse space-times, $R_{ij}V^iV^j$ corresponds to the energy density as observed by a timelike observer with tangent V^i, and our results show that this energy density diverges very powerfully in the vicinity of the naked singularity. This allows one to conclude that significant physical phenomena could take place in the vicinity of such a singularity which may not be regarded as a mere mathematical pathology. In fact, this strong curvature condition ensures that all the objects falling into the singularity are crushed to zero size and thus the space-time may not admit even any continuous extension through such a singularity. Thus, this turns out to be a genuine boundary point and actually the space-time extension through it may not make any sense as all volume forms along the trajectories falling to the singularity are reduced to zero size. The singularity in the Friedmann models at $t = 0$ and in the Schwarzschild space-time at $r = 0$ are such strong curvature singularities and the space-time does not admit any continuous extension through them.

An open issue that must be addressed is the question of stability of the naked singularity. If the naked singularity we have discussed ceases to exist under small perturbations, it may not be taken seriously. On the other hand, it is possible that just as the singularity theorems showed that the singularities forming in the spherically symmetric situation are stable under small perturbations from this symmetry, the naked singularity under consideration could also be stable in a similar manner. A detailed analysis is required of this issue, which of course needs a suitable criterion for stability to be formulated in general relativity. We discuss this issue in more detail in Section 7.4.

Subject to the reservations such as above, if we accept that naked singularities occur for a wide range of self-similar as well as non-self-similar space-times, the cosmic censorship requirement then implies that the forms of matter such as the dust, perfect fluid, or inflowing radiation must break down and cease to be good approximations as the gravitational collapse progresses to an advanced stage. In fact, as pointed out by Eardley (1987), the dust and null fluid may not be regarded as fundamental forms of matter even at the classical level but could be regarded as approximations to more basic entities such as a massive scalar field and a massless scalar field (in the eikonal approximation) respectively. Thus, the question could be asked whether naked singularities occur for a scalar field coupled to gravity, or for similar matter fields other than forms of matter such as dust, perfect fluid, or collapsing radiation.

To have an insight into this problem, we now generalize below the results of Section 6.6 on the self-similar collapse of a perfect fluid. It is pointed out below that the strong curvature naked singularities could occur in the self-similar gravitational collapse of any form of matter satisfying the weak energy condition for the positivity of mass-energy density. The relevance of such a question is that even though the form of matter such as a perfect fluid has a wide range of physical applications with the advantage of incorporating the pressure which could be important in the later stages of collapse, it is certainly important to examine if similar conclusions will hold for other reasonable forms of matter. The reason is, it is conceivable that the naked singularity would be an artifact of the approximation used, rather than being a basic feature of gravitational collapse. Thus, we need to examine the conclusions of self-similar collapse for a broader range of matter. This analysis shows that the conclusions of Section 6.6 are not limited to a specific form of matter such as a perfect fluid.

The non-zero metric components for a spherically symmetric space-time are $(t, r, \theta, \phi = 0, 1, 2, 3)$

$$g_{00} = -e^{2\nu}, \quad g_{11} = e^{2\psi} \equiv V + X^2 e^{2\nu}, \quad g_{33} = g_{22}\sin^2\theta = r^2 S^2 \sin^2\theta,$$

where V is defined as above and due to self-similarity ν, ψ, V, and S are functions of the similarity parameter $X = t/r$ only. The remaining freedom in the choice of coordinates r and t can be used to set the only off-diagonal term T_{01} of the energy momentum tensor T_{ij} to zero (using comoving coordinates). We assume the matter to satisfy the weak energy condition, that is,

$$T_{ij}V^iV^j \geq 0,$$

for all non-spacelike vectors V^i. Therefore, ${T_0}^0 \leq 0$ (that is, $T_{00} \geq 0$), ${T_1}^1 - {T_0}^0 \geq 0$, ${T_2}^2 - {T_0}^0 \geq 0$. The relevant field equations for a spherically symmetric self-similar collapse of the fluid under consideration can be written from Section 6.6 as follows:

$$\begin{aligned} {G^0}_0 = \frac{-1}{S^2} &+ \frac{2e^{-2\psi}}{S}\left(X^2\ddot{S} - X^2\dot{S}\dot{\psi} + XS\dot{\psi} + \frac{(S - X\dot{S})^2}{2S}\right) \\ &- \frac{2e^{-2\nu}}{S}\left(\dot{S}\dot{\psi} + \frac{\dot{S}^2}{2S}\right) = 8\pi r^2 {T^0}_0, \end{aligned}$$

$$\begin{aligned} {G^1}_1 = \frac{-1}{S^2} &- \frac{2e^{-2\nu}}{S}\left(\ddot{S} - \dot{S}\dot{\nu} + \frac{\dot{S}^2}{2S}\right) \\ &+ \frac{2e^{-2\psi}}{S}\left[-SX\dot{\nu} + X^2\dot{S}\dot{\nu} + \frac{(S - X\dot{S})^2}{2S}\right] = 8\pi r^2 {T^1}_1, \end{aligned}$$

$${G^2}_2 = 8\pi T_2^2, \quad {G^0}_1 = \ddot{S} - \dot{S}\dot{\nu} - \dot{S}\dot{\psi} + \frac{S\dot{\psi}}{X} = {T^0}_1 = 0.$$

Using the last equation above, the first two equations can be combined to get

$$\dot{V}(X) = Xe^{2\nu}[H - 2],$$

where $(\cdot)$ is the derivative with respect to the similarity parameter $X = t/r$ and $H = H(X)$ is defined by

$$H = r^2 e^{2\psi}(T^1{}_1 - T_0{}^0).$$

For matter satisfying weak energy condition, it follows that $H(X) \geq 0$ for all X. Here we have four field equations with six unknowns. The remaining two equations come from the choice of form of matter one is dealing with. Using the field equations above and methods similar to Section 6.6 one can see that the singularity at $t = 0, r = 0$ is naked when the equation $V(X) = 0$ has a real simple positive root, that is,

$$V(X_0) = 0,$$

for some $X = X_0$. We note that such a condition was used by Waugh and Lake (1989), and by Ori and Piran (1990) for some numerical simulations of self-similar gravitational collapse to examine the formation of naked singularity. Further, the relevance of the same for naked singularity formation in self-similar radiation collapse models was examined by Dwivedi and Joshi (1989, 1991). It is also seen that as in Section 6.6, a non-zero measure of future directed non-spacelike trajectories will escape from the singularity provided

$$0 < H_0 = H(X_0) < \infty.$$

Using the equations of non-spacelike geodesics in a self-similar spacetime, it is seen that these escaping trajectories near the naked singularity are given by

$$r = D(X - X_0)^{\frac{2}{H_0 - 2}}.$$

Here D is a constant labelling different integral curves, which are the solutions of the geodesic equations, coming out of the naked singularity. It is seen that $H_0 > 0$ will hold if the weak energy condition above is satisfied and when the energy density as measured by any timelike observer is positive in the collapsing region near the singularity. In this case, when $H_0 < \infty$, families of future directed non-spacelike geodesics will come out, terminating at the naked singularity in the past. On the other hand, for $H_0 = \infty$, a single non-spacelike trajectory will come out of the naked singularity. This characterizes the formation of naked singularity in self-similar gravitational collapse. Such a singularity will be at least locally naked and considerations such as those in Section 6.6 can be used to show that it could be globally naked as well provided $V(X_0) = 0$ has more than one real simple positive roots.

We do not go into a detailed discussion of the sufficient conditions which would ensure the real positive roots for $V(X_0) = 0$, and hence the existence of naked singularity, in terms of physical parameters involved and the initial data specified prior to the onset of the singularity. However, the existence of several classes of self-similar solutions to Einstein equations which we have discussed in detail in Chapter 6, where the gravitational collapse from a regular initial data results into a naked singularity, indicates that such a condition will be realized for a wide variety of self-similar collapse scenarios. For example, the cases of radiation collapse models with a linear mass function discussed in Section 6.5, and also the self-similar Tolman–Bondi models are all special cases of the treatment given here. For the case of radiation collapse with the mass function $m(u) = \lambda u$ (where u is the advanced time), the above initial condition corresponds to a restriction on the parameter λ (which is the rate of collapse) given by $0 < \lambda \leq 1/8$. The collapse must result into a naked singularity when λ is in this range, and an event horizon covers the singularity otherwise. Similar restrictions are obtained, ensuring the existence of a real positive root for $V(X) = 0$, for classes of Tolman–Bondi dust models as discussed in Section 6.8, and for the case of adiabatic perfect fluid in which case this turns out to be a fourth order algebraic equation. In general, for a given form of matter, the existence of a real positive root of $V(X_0) = 0$ will put a restriction on the range of values of physical parameters involved. One could treat this condition for the existence of a naked singularity as an initial value problem for the first order differential equation given above governing $V(X)$. That is, for a given form of matter one could solve this first order differential equation, with the initial value $V(X_0) = 0$, for some real positive value $X = X_0$. It is possible that the regular initial data specified for a realistic gravitational collapse might just always avoid this initial condition and could result into a black hole. On the other hand, all the initial data sets satisfying this condition will necessarily evolve into a naked singularity regardless of the form of matter involved. It is thus seen that a naked singularity would form in the collapse of a wide range of matter forms satisfying the weak energy condition.

Having shown the existence of non-zero measure of non-spacelike families coming out from the naked singularity in this general case, we could also examine the curvature strength of naked singularity which provides an important test of its physical significance. This was calculated in Section 6.6 along all the non-spacelike geodesics terminating at the naked singularity in the past to show that this is a strong curvature singularity in a powerful sense. Even for the case of the general form of matter considered here, this turns out to be a strong curvature naked singularity. We consider here only the radial null geodesics coming out, which are given by,

$$\frac{dt}{dr} = e^{\psi-\nu}.$$

The curvature strength of the singularity is measured by evaluating the limit of $k^2 R_{ij} V^i V^j$ along these trajectories near the singularity, which is given in the present case by

$$\lim_{k \to 0} k^2 R_{ij} V^i V^j = \frac{4H_0}{(2+H_0)^2} > 0.$$

It follows that this is a strong curvature naked singularity in the sense that the volume forms defined by all possible Jacobi vector fields vanish in the limit of approach to the naked singularity, in which case the space-time may not admit any continuous extension.

Though the above provides a general result on the occurrence of naked singularities in self-similar collapse, in order to have a detailed understanding of the nature and structure of the same it would be necessary to consider examples with specific equations of state and various different reasonable forms of matter. In this connection, we note the possibility indicated by Penrose (1974a), that according to the Hagerdorn equation of state (Hagerdorn, 1968) used in cosmology to describe the high density behaviour, the energy-momentum tensor becomes more and more dustlike as the densities grow higher and higher. The point is, there are considerable uncertainties about the equation of state at higher and higher densities and the form of matter that could be used as a good approximation. It follows that in order to rule out naked singularity and to generate a provable formulation of censorship hypothesis, this issue of reasonable forms of matter and equations of state at higher densities requires careful consideration. Without such an effort no progress in the direction of ruling out naked singularities appears possible.

A possibility is suggested in this direction by the above results on self-similar collapse. It follows from the considerations here and in Section 6.6 that if the weak energy condition requirements indicated above are violated then even though a naked singularity may exist (in the sense that a single null geodesic might escape from the singularity), no families of non-spacelike trajectories may terminate at the naked singularity in the past. Further, the curvature strength criterion also will not be satisfied in this situation. For all practical purposes such a naked singularity may not be taken seriously and the cosmic censorship hypothesis may be taken to be valid. What is required then is violation of the energy conditions in advanced stages of gravitational collapse. It is not clear what mechanism may be invoked to achieve this purpose, except probably the quantum effects which may become important in the very final stages of collapse, and the quantum gravity corrections which would be relevant at extremely small lengths of the order of the Planck scales. However, it must be noted that in such a case this no longer remains a purely classical formulation of the censorship hypothesis in a classical standard theory of gravitation such as the general theory of relativity.

It is clear from our considerations so far that the final fate of gravitational collapse remains an issue involving exciting possibilities. The existence of naked singularities in nature would be a matter for concern since they could radiate arbitrarily, giving rise to unpredictable astrophysical phenomena. Even if a successful quantum theory of gravity was somehow to dispense with all the space-time singularities, one would nonetheless have to take account of essentially singular, high density regions of matter and radiation predicted by the classical theory (even though the final singularity could be possibly avoided by quantum effects). Such regions will have properties dominated by quantum mechanical uncertainties and will thus, if visible to external observers, give rise to global predictability problems as before. This is the main reason why it is important to establish whether classical general relativity conforms to any cosmic censorship principle which constrains or prohibits the occurrence of naked singularities.

In fact, the status of naked singularity and cosmic censorship violation in quantum gravity remains an interesting open question. Particle creation near a shell crossing type of naked singularity was examined by Ford and Parker (1978), who did not discover a large flux of particles. However, a calculation with a similar purpose by Hiscock, Williams and Eardley (1982), for a shell-focusing singularity in Vaidya space-times with a linear mass function, reported that the energy flux of the created particles along the Cauchy horizon diverges. This result suggested that the back reaction on the metric from this flux of created particles might prevent the formation of naked singularity and censorship violation. On the other hand, Hawking (1992) has concluded that evaporation of two-dimensional black holes in the semi-classical approximation to quantum gravity necessarily gives rise to naked singularities, or else the semi-classical approximation breaks down.

What appears necessary at this stage is deducing general constraints on the naked singularities possible within the framework of general relativity. This would throw further light on their nature and structure and this is also necessary in order to arrive at any suitable version of the censorship hypothesis. We discuss some of these issues in the following sections.

7.2 Causality constraints

In this section we briefly consider the question of how a naked singularity affects the causal structure of space-time, and what constraints are imposed on the structure of naked singularity from the point of view of the causal structure of space-time.

A naked singularities influences the structure of space-time in two ways. On the one hand, its gravitational field affects the motion of surrounding matter. On the other hand, it disrupts the causal structure of space-time which in turn influences interaction processes. One may conjecture that a genuine naked singularity in nature must exhibit both these features to excess (Newman and Joshi, 1988). The naked singularity must have a strong gravitational field and it must also seriously affect the causal structure of space-time. The question then arises as to whether or not these two requirements are compatible. In fact, several earlier results of Newman (see for example, Newman, 1986b and references therein) and considerations such as Królak (1983) could be interpreted as providing constraints on the causal structure of space-time in the presence of naked singularities in the space-time.

To consider these issues in some detail in order to deduce certain general constraints on the structure of naked singularities, we work within the framework of space-times which are weakly asymptotically simple and empty in the sense defined in Chapter 4. These space-times are ideally suited to model collapse scenarios of isolated objects. The basic notion of weak cosmic censorship is that, in such a space-time, the data on some suitable partial Cauchy surface should determine the space-time geometry all the way out to the future null infinity $\mathcal{I}^+$ of the conformal completion.

Consider a Tolman–Bondi space-time (M, g) of Section 6.7 in which an inhomogeneous dust cloud collapses to form a globally naked central singularity of shell-focusing type as shown in Fig. 52. The conformal completion $(\overline{M}, \bar{g})$ here is partially future asymptotically predictable, but not future asymptotically predictable, from a partial Cauchy surface $\mathcal{S}$ of (M, g). The naked singularity may be regarded as causally innocuous in that it fails to obscure any space-time points from observers at $\mathcal{I}^+$. However, the gravitational field is sufficiently strong that, along each radial null geodesic falling into the naked singularity, all irrotational congruences of Jacobi fields are forced to reconverge, or in other words the strong curvature condition is satisfied. The limiting focusing condition of Newman (1986a) is also satisfied along all these null trajectories. The model contains no naked points at infinity in the sense that, for each compact set $\mathcal{K}$, every future endless, future complete null geodesic of (M, g) in $\overline{I^-(\mathcal{K}, \overline{M})}$ has a future endpoint in $\mathcal{I}^+$. On the other hand, for the space-time of Yodzis, Seifert and Muller Zum Hagen (1973), shown in Fig. 53, crossing shells of dust give rise to a globally naked singularity. Here again the conformal completion is partially future asymptotically predictable, but not future asymptotically predictable from the partial Cauchy surface. Here the naked singularity is gravitationally weak in that radial null geodesics falling into it are not subject to the limiting focusing condition. However, the causal disruption is such that the

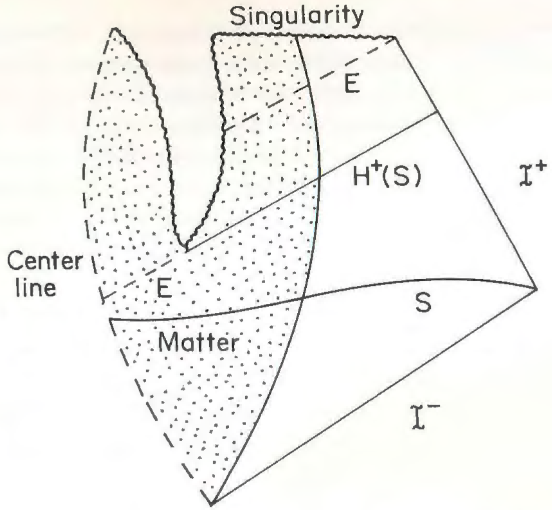

Fig. 53 The formation of a shell-crossing naked singularity in the inhomogeneous collapse of dust in Tolman–Bondi models. The event horizon E and the Cauchy horizon $H^+(S)$ are as indicated.

naked singularity obscures a region of space-time from observers at future null infinity. This corresponds to the fact that the even horizon intersects the domain of dependence of the partial Cauchy surface. There are no naked points at infinity in the space-time.

It was shown by Newman and Joshi (1988) that these examples demonstrate a genuine feature of naked singular space-times. In particular, when suitably interpreted, their result implies that the evolution of non-singular initial data on a suitable partial Cauchy surface cannot give rise to globally naked singularities which are both gravitationally strong, in the sense of refocusing of Jacobi fields, and severely causally disruptive in the sense of obscuring a region of space-time from observers at future null infinity.

Consider a space-time (M, g) admitting a weakly asymptotically simple and empty conformal completion $(\overline{M}, \bar{g})$, which is partially future asymptotically predictable from a partial Cauchy surface S of (M, g). The violation of future asymptotic predictability from S means only that at least one generator of future null infinity left the closure of the future domain of dependence of S in the conformal completion in future direction; however, in the present case we assume that this happens for every null generator. This means that if a naked singularity exists, it can be seen in every direction. Naked points at infinity are again assumed to be absent. Since it is conceivable that naked singularities may be associated with phenomena such as closed timelike curves (see for example, Clarke and de Felice, 1984), no causality conditions other than those implied by the existence of a partial Cauchy surface, are assumed. The condition that singularities are gravitationally strong is characterized by a requirement that future endless future incomplete null geodesics of the space-time are subject to the limiting focusing condition. A condition that naked singularities resulting from the data on $\mathcal{S}$ prevent a region of the space-time from being seen from the future null infinity is characterized by a requirement that the domain of dependence of $\mathcal{S}$ contains part of the event horizon.

A characteristic of the spherical symmetry of the Tolman–Bondi space-times of Fig. 52 and Fig. 53, depicting shell-focusing and shell-crossing naked singularities, is that the Cauchy horizon $H^+(\mathcal{S}, M)$ is devoid of caustics. In the following theorem (for the proof and further details we refer to Newman and Joshi, 1988) a limited deviation from the spherical symmetry is allowed with a condition that the caustics of this Cauchy horizon, if any, do not have an accumulation point on the future null infinity. Such a result can be interpreted as comparing the strength and causal structure disruptions caused by a naked singularity in the space-time.

Theorem 7.1. Let (M, g) be a space-time satisfying the null convergence condition $R_{ij}V^iV^j > 0$ for all null vectors V^i. Suppose the space-time admits a weakly asymptotically simple and empty conformal completion $(\overline{M}, \bar{g})$ and also a connected partial Cauchy surface $\mathcal{S}$ such that the following conditions are satisfied:

(a) $(\overline{M}, \bar{g})$ is partially future asymptotically predictable from $\mathcal{S}$,
(b) $(\overline{M}, \bar{g})$ violates weak cosmic censorship in the sense that every generator of $\mathcal{I}^+$ leaves $\bar{D}^+(\mathcal{S}, \bar{M})$ in the future direction,
(c) $(\overline{M}, \bar{g})$ admits no naked points at infinity,
(d) no point of $\mathcal{I}^+$ is an accumulation point of future endpoints in M of generators of $H^+(\mathcal{S}, M)$.

Then the following both cannot hold simultaneously:

(1) every future endless, future incomplete null geodesic of (M, g) in the set $D^+(\mathcal{S}, M)$ is subject to the limiting focusing condition;
(2) the event horizon of (M, g) has non-empty intersection with $D(\mathcal{S}, M)$.

It is also conceivable that closed timelike curves might arise in the vicinity of a naked singularity. For example, Clarke and de Felice (1984) considered the possibility of strong causality violation due to a naked singularity which could extend to the future null infinity. Also, for the Kerr space-time, when the rotation is sufficiently high, the space-time contains a naked singularity and also closed timelike curves. However, the connection between these two phenomena does not appear to be clear as yet and needs further investigation. The converse here is of course true in the sense that the existence of closed timelike curves would violate the strong cosmic censorship hypothesis. This is in the sense that the space-time would contain a Cauchy horizon and is not globally hyperbolic. In fact, Morris, Thorne and Yurtsever (1988) have pointed out that the issues of causality violation and cosmic censorship may be connected in the above sense. They show that the wormhole space-times considered by them admit closed timelike lines and the resulting Cauchy horizon is stable in a certain sense. It may be noted, however, that these wormhole space-times involve the violation of averaged weak energy condition (in the sense that there are null geodesics with the tangent vector V^i, passing through the wormhole along which $\int_0^\infty T_{ij}V^iV^j < 0$). Could one allow the violation of energy conditions in the above sense? One could probably answer this question only within a suitable framework of the quantum field theory and quantum gravity.

7.3 Cosmic censorship and topology change

The question as to whether the topology of a spacelike hypersurface can change as it evolves in time, and if so, what the implications are of such a phenomenon, has been of interest for classical general relativity as well as various possible quantum gravity frameworks. Even though the task of constructing a full quantum gravity theory is far from being complete at the moment, the role of topology changes or non-trivial space-time topologies in possible quantization schemes has been stressed by many authors (see for example, Wheeler (1964) and Friedmann (1991) and references therein). At the classical level it was pointed out by Geroch (1967) that if the space-time is causal and admits a Cauchy surface S, then all Cauchy surfaces are homeomorphic and hence no topology change can take place,the space-time as a whole being $S \times \mathbb{R}$. It is possible to relate topology change in space-times with certain formulations of cosmic censorship hypothesis at this level itself. The strong cosmic censorship formulation by Penrose (1979) states that departures from global hyperbolicity cannot occur in a space-time and hence it must admit a Cauchy surface which ensures a complete predictability in M. Thus, for a causal space-time, strong cosmic censorship ensures that no topology change can take place.

An interesting question that remains open is that of the connection between topology change in space-time and various other forms of cosmic censorship in general, and in particular, the weak censorship hypothesis within the framework of asymptotically flat space-times. This was examined by Joshi and Saraykar (1987) to show that topology change gives rise to singularities which are naked in the sense of those occurring in the shell-crossing models due to Yodzis, Seifert and Muller Zum Hagen (1973) or as

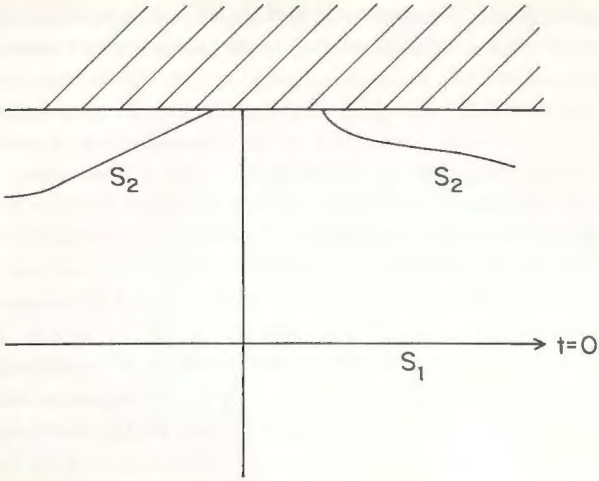

Fig. 54 The spacelike surfaces S_1 and S_2 are not homeomorphic; however, there is no topology change in the space-time.

defined by Newman (1984). It thus seems that in general topology changes are intimately related with the violation of cosmic censorship. We discuss below this result in some detail.

Firstly, we define the notion of topology change in a space-time. Consider an evolving sequence of spacelike hypersurfaces $S_1, S_2, ..., S_n$ with S_{i+1} contained in $I^+(S_i)$, where the topology of some S_i and S_{i+1} are not the same. However, such a topology change need not be genuine if the hypersurfaces chosen are not appropriate. For example, take two-dimensional Minkowski space-time with the region $t \geq 1$ amputed. Let S_1 be the $t = 0$ axis, where as S_2 may consist of any two disjoint spacelike pieces running into the cut but always remaining to the future of S_1. Clearly, S_1 and S_2 are not homeomorphic; however, there is no topology change because $t = 0$ serves as a Cauchy surface in the space-time (Fig. 54).

This difficulty can be cured by using Gowdy's notion of instantaneous Cauchy surfaces to characterize topology change in M (Gowdy, 1977). These are achronal surfaces whose Cauchy development interior is maximal on the set of all achronal surfaces. If now S and S' are two instantaneous Cauchy surfaces with $S' \subset I^+(S)$ and S and S' not homeomorphic, we say that topology change occurs in M. It is clear that if M admits topology change in any well-defined sense then it cannot be globally hyperbolic and hence must have a non-trivial Cauchy horizon. It is intuitively clear that this Cauchy horizon must be related in some sense with the existence of singularities in the space-time. However, this does not tell us about the location or nature of these singularities. The formulation of topology change in terms of instantaneous Cauchy surfaces helps in this respect as shown by the following result of Gowdy (1977). Suppose the space-time contains two non-homeomorphic instantaneous Cauchy surfaces S and S' with S acausal and $S' \subset I^+(S)$. Then the future Cauchy horizon $H^+(S)$ of S is non-empty and $I^+(S)$ is not globally hyperbolic. We note that the horizon $H^+(S)$ arising, when S is not a Cauchy surface is non-trivial in nature, that is, M cannot be globally hyperbolic when $H^+(S) \neq \emptyset$.

We have to analyse these horizons which arise as a result of topology change in M. Let S be an instantaneous Cauchy surface in M which undergoes a topology change in the above sense and let $H^+(S)$ be the resulting horizon. As discussed in Chapter 4, $H^+(S)$ will be an achronal surface which is generated by null geodesics which are either endless in $H^+(S)$ in the past or they meet on edge(S).

To examine the singular nature of $H^+(S)$ or the existence of space-time singularities, one would generally need to assume an energy condition for the space-time, which would typically ensure positivity of energy (and hence a well-behaved focusing effect) and additionally that the trajectories do encounter some non-zero matter or radiation. One could at least demand the generic condition discussed in Chapter 5, or, in a cosmological framework which we discuss here in the beginning, one may encounter the matter or radiation almost everywhere in the past. For this purpose, in addition to the weak energy condition, we also assume the positivity condition discussed in Chapter 5, namely that along all past-directed null geodesics we must have

$$\lim_{s \to k} \inf T_{ij} K^i K^j > 0, \tag{7.1}$$

along the trajectory, where k is the limit approached by the affine parameter s in the past. The presence of microwave background radiation and the expansion of the universe imply the naturalness of such a condition, where the matter and radiation densities must be increasing in the past. However, we would like to remark that the possibilities of weakening the energy conditions needed for gravitational focusing effects have been studied in detail (see for example, Tipler, Clarke and Ellis, 1980, for a review), and it is possible to replace the above condition by other weaker conditions. We shall not explore such possibilities here.

To consider the nature of null generators of $H^+(S)$, let $p \in H^+(S)$ and γ be a null generator with future end point p. The generator γ, when extended in the past, is either past endless or it can have end points only on edge(S), which is empty in our case. Now, suppose γ is complete in the past. Since γ lies on the null surface $H^+(S)$, it cannot contain any pairs of conjugate points as discussed in Chapter 5, because any two such conjugate points will be timelike related, which violates the achronal nature of $H^+(S)$. The generator γ is a member of the congruence of null geodesics and conjugate points along γ correspond to the expansion θ of the congruence being infinite, which is governed by the equation (5.13). It can be shown that, when the energy condition stated above is satisfied, there will be infinitely many conjugate points in the interval $(0, \infty)$ along γ in the past which contradicts the achronal nature of $H^+(S)$. We thus deduce that the null generator γ must be incomplete in the past. The implication is that topology change in a space-time to the future of a spacelike hypersurface S induces singularities to occur towards the future of S.

Before discussing the nature of these singularities, we address the next important question that arises in the present context: when the topology change occurs and as a result the strong cosmic censorship is violated, whether or not the weak cosmic censorship is also violated. However, this relationship between topology change and weak cosmic censorship has to be examined within the framework of weakly asymptotically simple and empty space-times characterized by the existence of a map $\theta : M \to \overline{M}$ which is a conformal isometry from M into its conformal compactification $\overline{M}$ in which infinity is represented by a null boundary $\mathcal{I}$. Here clearly eqn (7.1) does not hold, since it will certainly be violated on null geodesics ending on past null infinity $\mathcal{I}^-$. Such geodesics will be complete, which suggests that when we are considering weakly asymptotically simple and empty space-times we impose instead of eqn (7.1) a more general condition given as follows:

If γ is a complete past directed null geodesic in M with tangent vector K^i then either eqn (7.1) is satisfied or $\theta \circ \gamma$ must have a past end point on $\mathcal{I}^-$.

The preceding argument can now be adapted to show that, if S and S' are non-homeomorphic instantaneous Cauchy surfaces with $S' \subset I^+(S)$, and if the space-time is weakly asymptotically simple and empty, satisfying the condition above, then there exists an incomplete past directed null geodesic to the future of S.

To see this, first we note that as before, $H^+(S)$ is non-empty and we consider a generator γ of this null hypersurface. If γ is incomplete then nothing remains to be proved. If it is complete and γ does not end on $\mathcal{I}^-$ then the condition above reduces to the original condition eqn (7.1) and our previous argument applies to show that this is incompatible with completeness. Hence, one has only to eliminate the possibility that γ does end at some point p on $\mathcal{I}^-$. In such a case we must have $p \in \overline{\theta(S)}$. For if this were not so, we could find a neighbourhood U of p not intersecting

$\theta(S)$; points on $\theta \circ \gamma$ in U can be joined to $\mathcal{I}^-$ by a timelike curve, and the inverse image under θ of such a curve would provide an inextendible past directed timelike curve from points of $D^+(S)$ not meeting S, contradicting the definition of $D^+(S)$.

One could now derive a contradiction with the definition of an instantaneous Cauchy surface by proving that S can be deformed to give a spacelike surface with a larger domain of dependence. To this end, use in a neighbourhood U of p a null coordinate system $\{x^i\}$ with

$$\mathcal{I}^- \cap U = x^0 = 0, \quad \frac{D^2}{\partial s^2}[dx^0, dx^0] = 0, \tag{7.2}$$

and further,

$$g(\partial/\partial x^1, \partial/\partial x^i) = 0, \quad i = 1, 2, 3, \tag{7.3}$$

with $g(\partial/\partial x^1, \partial/\partial x^0) < 0$. Choose now a non-negative C^∞ function ξ with support in $[-1, 1]$, which is equal to 1 at 0 with $\xi'(x) \leq 0$ for $x > 0$, and set

$$u = (1 - x^0/\epsilon)\xi(x^0/\epsilon)\xi(x^0/\epsilon)\xi(x^0/\epsilon)\partial/\partial x^1, \tag{7.4}$$

where ϵ is a (small) parameter. Let $\phi(t)$ denote a general member of the one-parameter family of diffeomorphisms generated by u and set $S_t = \phi_t(S)$, which is well defined for small t and for ϵ so small that the deformation is confined to U. It can be shown using this function (Joshi and Saraykar, 1987) that the surfaces S_t remain spacelike for small enough values of t. It only remains to be shown finally that the $D(S_t)$ strictly contains $D(S)$, for small non-zero values of t. The argument used above to show that $p \in \overline{\theta}(S)$ shows that $\phi_t(p)$ does not belong to $D^+(S)$; so the required assertion will follow if we show that $S \subset D^-(S_t)$, which implies that $D(S) \subset D(S_t)$. But this follows directly from the construction of S_t, since curves can only leave the region between S_t and S via $\mathcal{I}^-$, which is not possible for future directed timelike curves. This gives a contradiction, excluding the case where $\theta \circ \gamma$ ends on $\mathcal{I}^-$ proving the result.

We point out that these singularities, arising as a result of topology change, will be naked in certain situations in the following sense as characterized by Newman (1984). The space-time M is called nakedly singular with respect to a partial Cauchy surface S when there exists $p \in H^+(S)$ such that $\overline{I^+(p)} \cap S$ is compact and the null generator γ of $H^+(S)$ through p is past incomplete. This definition is a generalization of the situation characterizing the shell crossing naked singularities discussed above where M admits a class of partial Cauchy surfaces relative to which it is future nakedly singular. In our situation, if we choose S to be compact, then for any $p \in H^+(S)$ the set $\overline{I^+(p)} \cap S$ will be compact and if γ is any incomplete generator through p, M will contain a naked singularity in the above sense. In any case, it is obvious intuitively that the singularities under consideration are naked in the sense that for the incomplete generator γ, $I^+(\gamma) \subset I^+(S)$, that is, the entire future of the singular generator γ is completely contained within the future of the instantaneous Cauchy surface S.

7.4 Stability

The cosmic censorship hypothesis, as formulated by Penrose (1969, 1979) emphasizes that the criteria of stability has to be satisfied in some sense, by requiring that space-times which are stable with respect to the changes in the initial data or the equation of state admit no naked singularities. Thus, for example, the strong censorship hypothesis can be stated as saying that stable space-times must be globally hyperbolic, or that they do not admit locally naked singularities. A similar statement for the weak censorship would be that stable space-times do not admit globally naked singularity. Hence, one would like to know if the gravitational collapse scenarios discussed in the previous chapter, and the naked singularities forming are stable in a suitable sense, say under small departures from the spherical symmetry or changes in the equation of state.

A general analysis on the question of stability is, however, a rather complicated issue because the stability theory in general relativity is a largely uncharted domain on which little is known. For example, an implication of strong censorship principle as stated above would be, singularities which are spacelike in nature must remain spacelike after a small perturbation in the space-time. One has to be somewhat careful, however, in formulating such a statement. The reason is, the Schwarzschild singularity could be thought of as being unstable in this sense, because the addition of even a small amount of charge or angular momentum changes the character of the singularity. In this case the solution changes to the Reissner–Nordström or Kerr space-time where the singularity is not necessarily spacelike. The addition of a small charge changes the model so that the singularity is locally naked. It is only when a *generic* type of perturbation is introduced in the space-time that one may hope the spacelike nature of the singularity to be retained. It is not clear, however, what such a generic perturbation would mean in general and one needs a well-defined stability theory in the framework of general relativity.

A possible approach in this connection, as indicated in Section 6.3, is to examine the stability of the Cauchy horizons which must form when the strong cosmic censorship is violated, that is, whenever global hyperbolicity breaks down. The Reissner–Nordström case provides an important clue here. In this case, the future Cauchy horizon extends all the way to spatial infinity and one way of specifying this is the following: Given any partial Cauchy surface S, the Cauchy horizon $H^+(S)$ associated with it has the property that given any point $p \in H^+(S)$, the set $\overline{I^-(p)} \cap S$ is non-compact. This is a different type of strong censorship violation as compared to that occurring in the Vaidya or the Tolman–Bondi case, where this set is always compact for any $p \in H^+(S)$. In such a case even a small perturbation in the initial data on the partial Cauchy surface will grow and diverge at $H^+(S)$ causing a blue-shift instability. The reason for this is the signals from far away regions on S are infinitely blue-shifted at $H^+(S)$, and in fact weak field perturbations would diverge along $H^+(S)$ (Simpson and Penrose, 1973; McNamara, 1978; Chandrasekhar and Hartle, 1983). In such a situation, what might occur really would be a curvature singularity rather than a Cauchy horizon, implying that $H^+(S)$ is not stable against sufficiently small perturbations in the initial data. This analysis of Reissner–Nordström case provides a strong evidence for the possibility that the Cauchy horizons, whenever they form should be unstable. In fact, such an instability is seen to be occurring for the wider class of space-times of the Kerr–Newman family, of which the Reissner–Nordström situation discussed above is a special case.

This form of instability of the Cauchy horizon is similar to that occurring in the Misner–Taub closed homogeneous cosmological models (Misner and Taub, 1969) which extend beyond a non-singular Cauchy horizon to a region which includes closed timelike curves. Misner and Taub examined the stability of this horizon and showed that this will be destroyed by small perturbations in the space-time. They showed the existence of photons which would be infinitely blue-shifted as they approach the Cauchy horizon and in such a case, the back reaction on the metric caused by such a blue-shift will completely destroy the Cauchy horizon. This can be illustrated by the example of the two-dimensional Misner space discussed in Chapter 4, which is obtained through a suitable identification in the two-dimensional Minkowski space-time. In this case the rightward travelling waves get boosted in frequency by a factor $(1+v)^{1/2}/(1-v)^{1/2}; v > 0$ with each passage through the line of identification. The waves must make this travel infinitely many times as they approach the Cauchy horizon and in the process the stress-energy for them has to diverge at this horizon. This instability might give rise to a curvature singularity destroying the Cauchy horizon and would presumably stop the evolution of the space-time into the region containing the closed timelike curves. This is an instability similar to that arising in the Taub-NUT space-time (Taub, 1951; Newman, Tamburino and Unti, 1963) at the Misner boundary (Fig. 55).

These examples are quite suggestive and prompt one to look for a general proof for the strong cosmic censorship in the form of instability

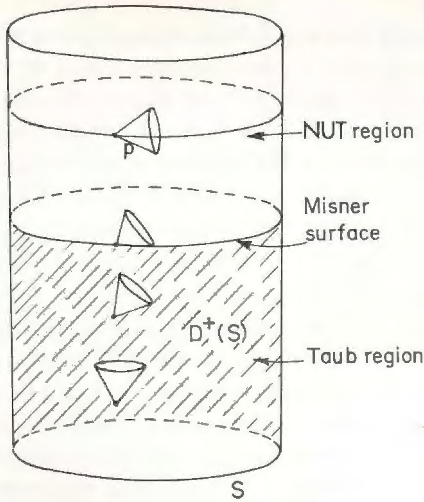

Fig. 55 The positions of the light cones are shown in the Taub-NUT space-time schematically. The Misner surface is the boundary of the domain of dependence for the partial Cauchy surface S, after which closed timelike curves occur in the space-time.

of Cauchy horizons whenever they form. This would establish that the violation of the strong censorship could never occur in a stable manner. Unfortunately, the situation again is not quite clear here as shown by some recent studies. For example, Morris, Thorne and Yurtsever (1988) have examined the stability of Cauchy horizons forming in the wormhole space-times that they studied. They show that the Cauchy horizon is stable in their space-times against the blue-shift mode of instability and they conjecture that actually this horizon will be fully stable, constructing a counter-example to the strong censorship hypothesis. There are photons in their space-time which travel along nearly closed null geodesics in future and they traverse the wormhole many times. These are blue-shifted each time and this is the kind of effect one would expect to generate the blue-shift instability. However, what happens is the light spreads fast enough and the energy of the photons is reducing fast enough which avoids the divergence. In this space-time, though there are many closed timelike curves, the set of closed null geodesics have measure zero and consequently, the initial data for dust particles could have unique evolution. Actually, the expectation is that there would be a sub-class of space-times which would have a well-defined initial value problem in spite of containing closed timelike curves.

Even though the Misner–Taub study discussed above suggests the instability of the Cauchy horizons for the homogeneous cosmological models, the further work by Ellis and King (1974), and King (1975) on these models indicates that more investigation is needed here. This is because the behaviour of a scalar field turns out to be stable and the Cauchy horizon need not be unstable in all the cosmologies they examined. The situation is much more unclear when one allows for the inhomogeneities. In fact, Clarke (1975) pointed out that global hyperbolicity is very unlikely in an inhomogeneous gravitational collapse, where the singularities may not form simultaneously at all places.

Coming to the gravitational collapse space-times which we examined in Chapter 6, admitting the formation of a naked singularity, the indication is that the Cauchy horizon there could be stable against at least the blue-shift mode of instability in certain cases such as a self-similar gravitational collapse. For example, Waugh and Lake (1989) examined the stability of Cauchy horizon in such models by injecting massless fields along the future Cauchy horizon in the high frequency (eikonal) limit. The frequency shift of this radiation is given by

$$\frac{\nu_e}{\nu_o} = \frac{(u_i k^i)_e}{(u_i k^i)_o},$$

where ν_o and ν_e are the observed and emitted frequencies respectively, and u^i is the four-velocity vector tangent to observer or emitter. Choosing the observer and the emitter to be comoving, it is seen that for the self-similar form of the metric given in Section 6.5, one gets $\nu_e/\nu_o = (T_o/T_e)^c$. Then the Cauchy horizon is stable as long as $c \geq 0$, which is ensured by the weak energy condition. Lemos (1992) discussed this stability of Cauchy horizon against the blue-shift mode of instability for the self-similar Vaidya and Tolman–Bondi space-times.

As noted above, the important question as far as the violation of both weak or strong censorship is concerned is whether the naked singularities forming are stable under small perturbations in the spherical symmetry or changes in the equation of state. Firstly, as far as departures from spherical symmetry are concerned, the situation could turn out to be similar to that of the singularity theorems, where it was shown that the singularities forming in the spherical situations such as the Schwarzschild or Robertson–Walker models are stable under small fluctuations in this symmetry. Further, as we shall point out in the next section, there is some evidence available now that naked singularities form in the non-spherically symmetric collapse as well. Thus, abandoning spherical symmetry need not get rid of naked singularities. Secondly, as far as the changes in the equation of state are concerned, our treatment in the previous chapter shows that the important features related to the naked singularities (namely, their existence, emission of families of non-spacelike curves from the same, and the strength of the singularity along all the trajectories terminating at the naked singularity in the past) persist even when the radiation equation of state is changed to matter in the form of dust, and also when pressure is introduced to include a perfect fluid. These are reasonable equations of state used widely in modelling astrophysical processes. Further, as far as the self-similar collapse is concerned, as pointed out in Section 7.1, these features on existence and nature of naked singularities generalize to all forms of matter satisfying the weak energy condition for the positivity of energy density.

The models considered in Chapter 6 are all spherically symmetric, however, if we take into account the above possibility of stability under departures from spherical symmetry, this implies the occurrence of naked singularities in a wide variety of gravitational collapse models, and for a whole range of reasonable matter. Further, we have also shown that the occurrence of naked singularities persists even when departures from the geometric symmetry of self-similarity of the space-time are allowed both in the radiation as well as the dust case, where the occurrence of naked singularities is observed also for the non-self-similar gravitational collapse.

We have discussed the general class of self-similar models in Section 6.5. The self-similarity condition is characterized by the existence of a homothetic Killing vector field for the space-time. In the following, we show that the existence of naked singularity is stable under a perturbation of these space-times to those containing a conformal Killing vector only. Specifically, self-similarity is equivalent to the existence of a homothetic Killing vector field ξ such that the Lie derivative of g_{ij} satisfies

$$L_\xi g_{ij} = 2g_{ij}. \tag{7.5}$$

Now, we consider a class of space-times which no longer satisfies the above but has the more general symmetry

$$L_\xi g_{ij} = \Omega^2(x^a) g_{ij}, \tag{7.6}$$

where Ω is a C^2 function of the space-time coordinates x^a. The space-time now admits a conformal Killing vector field rather than a homothetic Killing field. The class of space-times under consideration is now specified as follows: Suppose (M, g_{ij}) is a self-similar space-time as in Section 6.5, which contains a naked singularity. Consider now a small perturbation of the metric g_{ij} defined by $g_{ij} \to \bar{g}_{ij} = \Omega^2 g_{ij}$, where Ω is a function of the space-time coordinates such that

$$|\bar{g}_{ij} - g_{ij}| < \epsilon, \tag{7.7}$$

everywhere in the space-time, where $\epsilon > 0$ is any small constant. In this case, the new space-time $(\overline{M}, \bar{g}_{ij})$ is no longer self-similar but admits a conformal Killing vector and satisfies eqn (7.6). We examine the stability of the naked singularity of a self-similar space-time under the perturbations as specified above.

As pointed out in Section 6.5, a self-similar spherically symmetric collapse of perfect fluid in comoving coordinates $(u^a = \delta^a_t)$ is described by

$$ds^2 = -e^{2\nu(X)}dt^2 + e^{2\psi(X)}dr^2 + r^2 S^2(X)(d\theta^2 + \sin^2\theta\, d\phi^2). \tag{7.8}$$

The parameter $X = t/r$ is the similarity parameter and all functions of physical interest are functions of X. Therefore ν, ψ, and S are functions of X only. We have shown in Section 6.5 that a wide range of spherically symmetric self-similar collapse of a perfect fluid would result in a strong curvature globally naked singularity, where a non-zero set of outgoing null geodesics are specified by their tangent vectors K^a. We assume that the original space-time given by eqn (7.8) does contain a naked singularity and that perturbations to this self-similar spherically symmetric space-time are of the form

$$ds^2 = e^{2U}\left(-e^{2\nu(X)}dt^2 + e^{2\psi(X)}dr^2 + r^2 S^2(X)(d\theta^2 + \sin^2\theta\, d\phi^2)\right). \tag{7.9}$$

Here $\Omega = e^U$ and $U = U(t, r, \theta, \phi)$ is a C^2 function describing the perturbation, which is small enough everywhere so that eqn (7.7) is satisfied throughout the space-time for some $\epsilon > 0$.

Let K^a and $\bar{K}^a$ be tangents to the null geodesics in the unperturbed and perturbed space-times respectively:

$$\bar{K}^a = \frac{dx^a}{d\bar{\lambda}} = e^{2U}, \qquad K^a = e^{2U}\frac{dx^a}{d\lambda}, \tag{7.10}$$

where λ and $\bar{\lambda}$ are affine parameters to the null geodesics in unperturbed and perturbed space-times and are related by

$$\bar{\lambda} = \int e^{-2U} d\lambda. \tag{7.11}$$

At the singularity $t = 0, r = 0$, both $\bar{\lambda} \to 0, \lambda \to 0$. The trajectories of the null geodesics in the (t, r) plane for the perturbed space-time are

$$\frac{dr}{dt} = \frac{\bar{K}^r}{\bar{K}^t} = \frac{K^r}{K^t}. \tag{7.12}$$

Since the null geodesics terminate at the singularity $t = 0, r = 0$ with a definite tangent $X = X_0$ (X_0 being the real positive root of equation (6.72) or (6.73)) in the unperturbed space-time and U being C^2 (with $\infty > e^{2U}_{r\to 0, t\to 0} > 0$), it follows that null geodesics would terminate at the singularity with the same value of tangent $X = X_0$ even in the perturbed space-time, which would have a globally naked singularity at $t = 0, r = 0$ if the original space-time has one.

In case of self-similar spherically symmetric space-time we have shown that the strong curvature condition is satisfied along all the non-spacelike curves terminating at the naked singularity. This means that in the limit of approach to singularity

$$\lim_{k\to 0} k^2 R^{ab} K_a K_b > 0. \tag{7.13}$$

We now determine whether after the perturbation the strong curvature condition still holds along all null geodesics that terminate at the singularity. The Ricci tensor $\bar{R}_{ab}$ for the perturbed space-time is given by

$$\bar{R}_{ab} = R_{ab} - 2\left(U_{,a;b} - U_{,a}U_{,b} + \frac{g_{ab}}{2}(U^{;a}_{;a} + 2U_{,a}U^{,a})\right). \tag{7.14}$$

Further, we have

$$\bar{T}_{ab} = T_{ab} + T^{(p)}_{ab} \equiv T_{ab} + 2\left(-U_{,a;b} + U_{,a}U_{,b} + g_{ab}\left(U^{;a}_{;a} + \frac{U_{,a}U^{,a}}{2}\right)\right), \tag{7.15}$$

where $\bar{T}_{ab}$, T_{ab}, and $T^{(p)}_{ab}$ represent total stress-energy tensor of the perturbed space-time, unperturbed space-time, and small perturbation due to the perturbation factor U respectively. We have therefore

$$\begin{aligned}\lim_{\bar{\lambda}\to 0}(\bar{\lambda}^2 \bar{R}_{ab}\bar{K}^a\bar{K}^b) &= \lim_{\lambda\to 0}\left(\left(\int e^{-2U}d\lambda\right)^2 e^{4U} R_{ab}K^aK^b\right) \\ &+ \lim_{\lambda\to 0}\left(\left(\int e^{-2U}d\lambda\right)^2 e^{4U}(-2\ddot{U} + \dot{U}^2)\right),\end{aligned} \tag{7.16}$$

where a dot represents the derivative with respect to the affine parameter λ. It follows from the above equations and the l'Hospitals rule that contribution of the second term in eqn (7.16) vanishes and we have

$$\lim_{\bar{\lambda}\to 0}\left(\bar{\lambda}^2 \bar{R}_{ab}\bar{K}^a\bar{K}^b\right) = \lim_{\lambda\to 0}\left(\left(\int e^{-2U}d\lambda\right)^2 e^{4U}R_{ab}K^aK^b\right) > 0. \quad (7.17)$$

Hence it turns out that perturbation does not effect the curvature strength of the naked singularity.

Another important question that remains is the nature of matter near the singularity, that is, whether the perturbations alter the physical reasonableness of matter in the sense of satisfying an energy condition near the singularity. This point is important because as shown earlier, if weak energy condition is not violated near the singular region than a non-zero set of optical pulses(null geodesics) escaped from the singularity, otherwise only an instantaneous exposure to the singularity will be there in the form of a single null geodesic. Secondly, if the introduction of a small perturbation makes the matter near the singularity physically unreasonable in the sense that energy conditions are violated, than one may not regard the persistent existence of naked singularity after the perturbation as very physical.

Let V^a be a unit timelike vector; then using eqn (7.15),

$$e^{2U}\bar{T}_{ab}V^aV^b = T_{ab}V^aV^b + 2((U_{,a}V^{,a})^2 - U_{,a;b}V^aV^b - \left(U_{;a}^{;a} + \frac{U_{,a}U^{,a}}{2}\right). \quad (7.18)$$

In case the perturbation is such that spherical symmetry is not altered (that is, $U = U(t,r)$), it is possible to see from the above considerations that near the singularity $r = 0, t = 0$ contributions from the perturbation terms involving U behave like $1/r$. Therefore, near the singularity we have

$$\bar{T}_{ab}V^aV^b \propto \frac{c}{r^2} - \frac{d}{r}, \quad (7.19)$$

where $c > 0$ and hence the regions near the singularity do satisfy the weak energy condition. However, if the perturbations distort the spherical symmetry that is, U is a function of θ, ϕ also, then approximating the above equations near the singularity one can see that energy conditions near the singularity would be satisfied as long as $U_{,\theta\theta} - U_{,\theta\phi} \geq 0$ and $U_{,\phi\phi} - U_{,\theta\phi} \geq 0$.

7.5 Non-spherical collapse and alternative conjectures

What will be the final fate of gravitational collapse which is not exactly spherically symmetric? As observed in Sections 6.1 and 6.2, the main phases of the spherically symmetric collapse of a homogeneous dust cloud are the instability, implosion of the matter and the subsequent formation of horizon and a space-time singularity of infinite density and curvatures and infinite gravitational tidal forces. The singularity in this case is completely hidden within the horizon and hence causally disconnected from the observer at infinity. We investigated the consequences of the situation when the collapse is no longer homogeneous, and of different equations of state for the matter, still within the spherically symmetric framework, in Chapter 6.

Small perturbations over the spherically symmetric situation were taken into account in the work of Doroshkevich, Zel'dovich and Novikov (1966), de la Cruz, Chase and Israel (1970), and in the perturbation calculations of Price (1972). The main outcome of these papers is that the basic result of the spherically symmetric collapse situation remains unchanged, at least in the sense that an event horizon will continue to form in the advanced stages of the collapse. The paper by Doroshkevich *et al.* indicated that as the collapse progresses and the star reaches the horizon, the perturbations remain small and there are no forces arising to destroy the horizon. This paper contained the basic idea of the no-hair theorems for black holes which were developed later, as it pointed out that from the point of view of an outside observer, non-spherical perturbations in the geometry and the electromagnetic field die down as the star approaches the horizon. These findings were supported by the numerical calculations of de la Cruz *et al.* and Price stated above, where linearly perturbed collapse models were integrated to show that the perturbations died out with time as far as they could be followed. It may be noted, however, as pointed out by Israel (1986a), that one would basically desire a result showing that a small change in the initial data perturbs the solution only slightly. The above works lead to such a result for bounded time intervals only, whereas the existence of horizon depends on the entire future behaviour of the solution.

The next question is, do horizons still form when the fluctuations from the spherical symmetry are high and the collapse is highly non-spherical? It is known, for example, that when there is no spherical symmetry, the collapse of infinite cylinders do give rise to naked singularities in general relativity, which are not covered by horizons (Thorne, 1972; Misner, Thorne and Wheeler, 1973). However, what one has here in mind are finite systems in an asymptotically flat space-time. Not much is known in this connection except the *hoop conjecture* of Thorne (1972), which characterizes the final fate of a non-spherical collapse as follows: The horizons of gravity form when and only when a mass M gets compacted in a region whose circumference in *every* direction obeys

$$\mathcal{C} \leq 2\pi(2GM/c^2). \tag{7.20}$$

Thus, unlike the cosmic censorship conjecture, the hoop conjecture no longer rules out *all* the naked singularities but only makes a definite assertion on the occurrence of the event horizons in gravitational collapse.

The hoop conjecture allows for the occurrence of naked singularities in general relativity when the collapse is sufficiently aspherical, especially when one or two dimensions are sufficiently larger than the others. A known example in this connection is the Lin, Mestel and Shu (1965) instability in Newtonian gravity (see for example, Thorne, 1972; Shapiro and Teukolsky, 1991 for a discussion). Here a non-rotating homogeneous spheroid collapses, maintaining its homogeneity and spheroidicity but its deformations grow. If the initial condition is that of a slightly oblate spheroid, the collapse results into a pancake singularity through which the evolution could proceed. However, for a slightly prolate spheroidal configuration, the matter collapses to a thin thread which ultimately results into a spindle singularity. This is more serious in nature in that the gravitational potential and force and the tidal forces blow up as opposed to only density blowing up in a merely shell-crossing singularity. Even in the case of an oblate collapse, the passing of matter through the pancake causes prolateness and subsequently a spindle singularity again results without the formation of any horizon.

It was indicated by the numerical calculation of Shapiro and Teukolsky (1991) that a similar situation maintains in general relativity as well in conformity with the hoop conjecture. They evolved collissionless gas spheroids in full general relativity which collapse in all cases to singularities. When the spheroid is sufficiently compact, a black hole forms which contains the singularity, but when the semimajor axis of the spheroid is sufficiently large, a spindle singularity results without an apparent horizon forming. One could treat these only as numerical results, as opposed to a full analytic treatment, which need not be in contradiction to a suitably formulated version of the cosmic censorship. However, this gives rise to the possibility of occurrence of naked singularities in the collapse of finite systems in asymptotically flat space-times, which could be in violation of weak cosmic censorship but in conformity with the hoop conjecture.

A somewhat broader statement in a similar spirit as the hoop conjecture is the *event horizon conjecture* of Israel (1984, 1986a,b). A broad statement for this conjecture is given by the requirement that an event horizon must form whenever a matter distribution (satisfying appropriate energy conditions) has passed a certain critical point in its gravitational collapse, namely the formation of a closed trapped surface. A strong motivation for believing in this conjecture is that, unlike cosmic censorship hypothesis, no counter-examples to the same have been found so far. For example, the inequality (discussed in Section 6.3),

$$A < 4\pi(2M)^2, \tag{7.21}$$

must hold if the event horizon conjecture is true, for the area A of a closed trapped surface which forms at time $t = 0$ (say) in the gravitational collapse of mass M. The inequality (7.21) depends only on the initial data on the spacelike surface $t = 0$ and hence there is no need to trace future evolution of the system in order to verify the same. The efforts by Penrose and Gibbons (1972) did not succeed in generating any counter-examples to the above inequality and Ludvigsen and Vickers (1983) have provided a proof of this inequality under fairly general conditions.

In the case of the exact spherical symmetry holding for the space-time, it is known (see for example, Leibovitz and Israel, 1970) that an event horizon must form when a star collapses to a sufficiently small radius, and when the positivity of energy is satisfied. The only alternative to this is the star must radiate away all of its mass in the process of collapse; this appears very difficult for the stars having tens of solar masses. Even for non-spherical collapse, which involves only small perturbations from the spherical symmetry, the numerical calculations mentioned above seem to indicate that a non-singular event horizon must develop during collapse. As noted by Jang and Wald (1977), without an event horizon to stop the flow of outgoing radiation, the gravitational mass of the star must become negative. In fact for all the classes of space-times discussed in Chapter 6 as well, an event horizon always forms in the process of collapse, even though it fails to cover the singularity completely. This results into the formation of naked singularities in the space-time. Of course, in such a case, when the event horizons form accompanied by the naked singularities, the usual interpretation of the black hole physics becomes blurred due to the back reaction on the metric as a result of the emission from the naked singularities. For example, as our detailed analysis for the case of radiation collapse shows, the horizon could be fully covered by the emissions from the strong curvature naked singularity, some of which may fall back inwards, thus affecting the area and structure of the horizon.

As pointed out by Israel (1984), there are two main aspects of the formulation of such an event horizon conjecture. Firstly, one has to characterize the formation of a trapped surface in terms of the initial data on a spacelike surface in the form of a statement such as a closed two-surface S will be trapped if the gravitational mass interior to S exceeds a certain critical value (defined in terms of geometry of S). In this connection, the works such as those of Schoen and Yau (1983) and Ludvigsen and Vickers (1983) would be relevant. Whereas this could be done for spherical symmetric case, the situation could be fairly complicated in general, because one would need to characterize the concept of 'mass interior to S' precisely. One would probably need a concept such as the quasi-local mass as defined by Penrose (1982) and also investigated by Horowitz and Strominger (1983). The other aspect is that of time evolution of such a trapped surface which should extend for a finite distance in future to generate a spacelike three-cylinder, the sections of which are trapped surfaces with bounded dimensions. Assuming energy conditions, one would desire the regularity of the energy momentum tensor on this three-cylinder. If these bounds and extension depend only on the initial geometry and the mass within the trapped surface, then, as pointed out by Israel (1986a,b), one could extend the three-cylinder infinitely into the future provided it encounters no singularities in its future development.

These ideas hold in the spherically symmetric case which again provides a good illustration for the same. One could write the metric in the form

$$ds^2 = -e^{2\nu(r,t)}dt^2 + e^{2\psi(r,t)}dr^2 + r^2 d\Omega^2. \tag{7.22}$$

Here $e^{2\psi(t,r)} = [1 - 2M(r,t)/r]^{-1}$ and $M(r,t)$ is identified with the gravitational mass within a radius r at time t as suggested by the Einstein field equations, given in this case by

$$\frac{\partial M}{\partial r} = -4\pi r^2 T^0{}_0, \qquad \frac{\partial M}{\partial t} = -4\pi r^2 T_{10}. \tag{7.23}$$

In this case, a sphere with $r < 2M(r,t)$ is trapped and its cylindrical extension with $r = \text{const.}$ is a spacelike hypersurface. This is because, as a consequence of the strong energy condition, $-T_{10} > 0$, the mass function $M(r,t)$ must be an increasing function of time.

8

GLOBAL UPPER LIMITS IN COSMOLOGY

In this chapter, we apply the global concepts developed in Chapters 4 and 5 to deduce certain general upper limits in cosmology. Specifically, our aim is to allow for perturbations from the assumed homogeneity of the universe and to deduce in this general scenario the upper bounds on the age of the universe as well as limits on the elementary particle masses which were produced copiously during the early phases of the universe and which may be filling the universe today.

The Friedmann–Robertson–Walker universes discussed in Section 3.6 are based on the assumption that the universe is exactly homogeneous and isotropic. These simplifying assumptions allow the Einstein equations to be completely solved and the evolution of the universe is fully determined in the form of spatially open and closed models. There is an all encompassing big bang singularity in the past in these models as the origin from which the universe emerges in a very hot phase and continues its expansion as it cools. The present epoch t_0 denotes the time elapsed from this singularity, or the age of the universe. The values of three parameters H_0, q_0 and ρ_0 fully characterize these models, which are called the *Hubble constant*, the *deceleration parameter*, and the *density parameter* respectively. In fact, using the Friedmann equations, it is seen that the knowledge of any two of these parameters, say H_0 and q_0, determine all the rest of the parameters, such as the age t_0 and ρ_0. It is effectively the present value of ρ_0 or q_0 which determines whether the universe is open or closed, and its future evolution. Considerable effort is expended in the observational cosmology today towards determining the values of the Hubble constant and the deceleration parameter. Whereas the Hubble constant is believed to be anywhere in the range of 50–100 km $s^{-1}Mpc^{-1}$, there is a considerable uncertainty in determining the value of deceleration parameter and as a result, the question of the present density of the universe, or whether it is an open or a closed model, remains open. The determination of present age and density of the universe are two very important issues in cosmology, as they determine the future evolution and the nature of the universe.

Another important aspect of the big bang models is that in the very early phases in the universe after the big bang, the temperatures were arbitrarily high. The description of matter and interactions at such high energies are governed by elementary particle physics. The processes that occurred in the early universe would leave relics which may remain in the universe as it expands and cools. For example, at temperatures $T \geq 10^{11}$ K, neutrinos will be produced and destroyed in the weak interactions such as $e^+e^- \leftrightarrow \nu\bar{\nu}$. As the temperatures decrease, weak interactions cross sections rapidly decrease and at lower temperatures such interactions are not likely. Then such remnant neutrinos form an inert background in the universe and will still survive today if they have sufficiently long lifetimes (see for example Dolgov and Zeldovich, 1981; Wilczek, 1991, for a review of basic ideas). Similarly, if the local super symmetry is spontaneously broken, then the gravitinos will be copiously produced in the early universe. In all cases, the density parameter ρ_0 would provide the upper limit on density of matter in all its forms in the universe which includes such particles.

As we point out in Section 8.1, subject to the uncertainties mentioned above, the values of H_0 and q_0 fix the present age t_0 and the present density ρ_0 in terms of H_0 and as a function of q_0 depending on the nature of model, that is, whether it is open or closed. Thus, the big bang models provide two very important inputs in cosmology and also for particle physics, namely the age of the universe and mass limits for elementary particles such as neutrinos produced in the early universe.

It must be noted, however, that the assumption of global homogeneity used in these models represents a very high degree of idealization which is not observationally verifiable in a direct manner. Even if valid, it is not clear at what scale such a homogeneity of matter distribution will be achieved in the universe. It is clear that locally there are inhomogeneities in the universe in the form of individual stars clustering to form galaxies. The galaxies are not uniformly spread in space but form clusters of galaxies, and such clusters again form super clusters. If such a hierarchy continued, the homogeneity will not be achieved for the mass distribution. In fact, such hierarchical clustering was indicated as a valid possibility by de Vaucouleurs (1970) when the universe need not achieve homogeneity at any scale. Again, recent observations on the structures in the universe (Geller and Huchara, 1989; Saunders *et al.* 1991) have shown that galaxies and their clusters form large-scale super structures in the universe of the order of 100 Mpc, separated by large voids in the space which are almost empty. It follows that if the homogeneity is achieved, it must be at a very large scale which is not precisely known presently.

Under the situation, it will be very useful to have a cosmological framework which takes into account perturbations over the exactly homogeneous Friedmann models and allows for inhomogeneous distribution of matter in the universe. The purpose of this chapter is to show that this can be achieved in the framework of globally hyperbolic space-times using certain global properties we have developed earlier. In Section 8.2 we show that general upper limits can be derived on the age of the universe, without assuming the exact homogeneity, by considering the gravitational focusing of matter on the particle trajectories in the past. In Section 8.3 we consider particle mass upper limits in this general cosmological scenario and show that several meaningful mass upper bounds are derived even without the homogeneity assumption. The limits on the cosmological constant values are considered in Section 8.4 and in Section 8.5 we again consider the Friedmann models to obtain certain improvements over the results obtained earlier on ages and upper limits within this exact framework. While the considerations here obtain several useful limits for cosmology, they also serve to illustrate the applications of global methods in this field, representing an important global aspect in cosmology.

8.1 Upper limits in Friedmann models

The Friedmann equations of Section 3.6 can be integrated to get the following expressions for the age of the universe at the present epoch (see, for example, Weinberg, 1972).

$$t_0 = H_0^{-1} f(q_0), \tag{8.1}$$

where the value of the function $f(q_0)$ is given for the different models of universe as

$$\begin{aligned}
f(q_0) &= (1-2q_0)^{-1} - q_0(1-2q_0)^{-3/2}\cosh^{-1}\left(\frac{1}{q_0}-1\right) \quad \text{for} \quad 0<q_0<\tfrac{1}{2},\\
f(q_0) &= \tfrac{2}{3} \quad \text{for} \quad q_0=\tfrac{1}{2},\\
f(q_0) &= \frac{q_0}{(2q_0-1)^{3/2}}\left[\cos^{-1}\left(\frac{1}{q_0}-1\right) - \frac{(2q_0-1)^{1/2}}{q_0}\right] \quad \text{for} \quad \tfrac{1}{2}<q_0<\infty.
\end{aligned} \tag{8.2}$$

These models correspond respectively to an open Friedmann model, the Einstein–de Sitter universe and a closed model which will recollapse into a future singularity again. The function $f(q_0)$ decreases monotonically with q_0 with a maximum value of unity at $q_0 = 0$ and approaches the value zero as $q_0 \to \infty$. It follows that in all cases the age of the universe is less than H_0^{-1}. Further, we note that the Einstein equations relate the measurable Hubble constant and the deceleration parameter values with the present value of the density parameter by the relation

$$\rho_0 = \frac{3H_0^2 q_0}{4\pi G}. \tag{8.3}$$

Clearly, there is a close relationship between the present value of q_0 and the age, because a higher value of the deceleration parameter means the expansion is decreasing rapidly. This means it was faster in the past and the value of the age will be smaller. If the universe is younger today than $\frac{2}{3}H_0^{-1}$ then it must ultimately start recontracting and collapse to the future singularity, but otherwise it will expand forever. There is a great degree of uncertainty over the measurements of the deceleration parameter but from the optical observations it appears safe to conclude that very probably the inequality $q_0 < 10$ is not violated.

As pointed out above, the present measurements on the expansion parameter H_0 leave an uncertainty of a factor of two. If the value of H_0 is near its upper end then the universe might turn out to be uncomfortably young even to accommodate the oldest objects such as the globular clusters. Another way of accommodating a high value of the Hubble constant is to have a non-zero cosmological constant as we will discuss in Section 8.4. However, if we desire a vanishing cosmological constant and still a high value of the density, called the *critical density* corresponding to the value $q_0 = 1/2$ as required by the inflationary scenario, then this *age problem* becomes particularly severe. This requirement of having a critical density for the universe appears consistent with the observed near flatness of the spatial geometry of the universe and is closely related to the issue of dark matter in the universe. We shall discuss this below briefly.

Tracing the universe back in time using the Friedmann equations, we have seen in Section 3.6 that the radiation density grows faster than the matter density. As a result the universe would be radiation dominated in its early phases. This radiation density will grow arbitrarily high near the singularity and thus arbitrarily high temperatures are realized in the vicinity of the big bang. The particle physics theories predict that at temperatures higher than 10^{11} K, neutrinos will be produced and destroyed in weak interactions. The simplest theories of unification of the strong and weak interactions based on $SU(5)$ symmetry would require the neutrino to be massless, but many other grand unified theories allow the neutrino to have a non-zero rest mass. Such massive neutrinos with a small rest mass, created by pair production in the early universe, would have survived until now, forming an inert background just like the microwave background.

The existence of such massive neutrinos or other massive particles such as the axion or gravitinos, which we shall discuss later, could be very crucial for the cosmology as they might help solving the issue of the dark matter in the universe. There is considerable evidence that large quantities of non-luminous matter is present in the universe. The orbital speeds of hydrogen clouds orbiting around the galaxies are quite at variance with the observed luminous matter in the galaxy. If most of the matter in the galaxy were concentrated in the optically bright region only, then the velocity of such clouds will vary with the radial distance as $V(r) \propto r^{-1/2}$. The observation is, however, that for most galaxies $V(r)$ becomes essentially constant, which indicates the presence of a dark matter component in the galactic mass distribution with a density profile given by $\rho \propto 1/r^2$. It also follows from applying the virial theorem to rich clusters of galaxies that they must contain about 10 to 50 times their visible mass as a dark matter component.

The possible nature of such dark matter and its composition is not clear at the moment but elementary particles such as neutrinos with a small mass provide a distinct possibility to supply the required non-luminous mass, if they are present in sufficient number density in the universe (see,for example, Sciama, 1990). Consider now the possibility that neutrinos with a small rest mass were produced abundantly in the early universe. After the decoupling of matter from the radiation as the universe cools and expands, they will expand in a homogeneous manner as an inert background together with the microwave radiation, while the ordinary matter starts forming inhomogeneities in the form of galaxies and their clusters and super clusters and even larger structures. There are three species of neutrinos considered, namely the electron, muon, and the tau neutrinos, and the present number density of each of these species of neutrinos will be $\frac{3}{4} \times \frac{4}{11}$ times the number density of the microwave photons, which turns out to be around 110 cm^{-3} for each neutrino type (Sciama, 1990). The factor of 3/4 comes from difference in statistics whereas the additional photon production by e^+e^- annihilation processes give rise to the factor of 4/11. Now, as indicated above, most of the matter content should be in the form of dark matter, in which case we could take the universe to be neutrino dominated. Then, as pointed out by Cowsik and MacClelland (1972), Szalay and Marx (1976), and Gershtein and Zel'dovich (1966), from the observed values of H_0 and q_0 one obtains the density parameter ρ_0 using eqn (8.3), which provides the upper limit to the neutrino density. Using the number density of neutrinos given above and summing over all the three species then gives the upper limit to $\sum_i m_i$, which turns out to be few tens of electronvolts (depending on the values of the H_0 and q_0 used from the allowed spectrum of uncertainties). So far it has not been possible to confirm experimentally that neutrinos do have a non-zero rest mass; however, various laboratory experiments have been able to place limits on neutrino masses and the above cosmological limits compare very favorably with these which are given by $m_{\nu_e} < 8\mathrm{eV}$, $m_{\nu_\mu} < 250\mathrm{keV}$, and $m_{\nu_\tau} < 35\mathrm{MeV}$ (Winter, 1990).

8.2 General upper bounds on age of the universe

We model here an inhomogeneous and anisotropic universe by means of a general globally hyperbolic space-time. As discussed earlier in Chapter

4, these are space-times which admit a spacelike hypersurface S, called a Cauchy surface. The initial data specified on such a hypersurface can be evolved into the future or past to predict the future or past states of the universe by means of hyperbolic differential equations. Geroch (1970a) showed that such a space-time can be covered by a one-parameter family S_t of spacelike Cauchy surfaces. Thus, the state of the universe at any given epoch t can be referred to in terms of the surface of cosmic simultaneity $t = \text{const}$. Well-known cosmological space-times such as the Friedmann–Robertson–Walker models and Bianchi, or steady-state cosmologies, are globally hyperbolic.

We would like to investigate within this globally hyperbolic framework the extension into the past of timelike geodesic trajectories by considering the gravitational focusing effect of the matter in a space-time. Let the present epoch be characterized by a spacelike global Cauchy surface S_0, where we set $t = 0$, and the matter distribution on S_0 is given by the stress-energy tensor T_{ij} which satisfies the strong energy condition that there are no negative energy fields in the space-time, and $(T_{ij} - \frac{1}{2}g_{ij}T)V^iV^j \geq 0$, where V^i is a unit timelike vector. Further, let the dynamics of the universe be governed by the Einstein equations, then the energy condition gives $R_{ij}V^iV^j \geq 0$ for all matter fields. Though we have not required the exact homogeneity or isotropy of mass-energy distribution on S_0, we assume, for the simplicity of consideration, that there exists a minimum for density distribution on S_0. In view of the observed expansion of the universe, this should exhibit a non-decreasing behaviour in the past. This means that there exists some $k > 0$ such that

$$R_{ij}V^iV^j \geq k > 0, \tag{8.4}$$

at the present and all past epochs.

The gravitational focusing effect of matter in a space-time can be characterized by the concept of a point conjugate to a spacelike hypersurface S_t along a timelike geodesic $\gamma(t)$, orthogonal to S_t. Consider a congruence of timelike geodesics orthogonal to S_t. Let $\gamma(t)$ be a member of the congruence, then a point q along $\gamma(t)$ is said to be conjugate to S_t if neighbouring timelike geodesics orthogonal to S_t intersect at q. Such a situation arises when the expansion θ of the congruence becomes infinite at q, which is governed by the Raychaudhuri eqn (5.13) of Chapter 5. A timelike geodesic $\gamma(t)$ will be orthogonal to S_0 provided the expansion θ along $\gamma(t)$ satisfies

$$\theta = \chi^i{}_i, \tag{8.5}$$

at S_0, where χ_{ij} is the second fundamental form of the spacelike surface. With the substitution $\theta = z^{-1}(dz/dt)$, with $z = x^3$, the Raychaudhuri equation for timelike geodesics can be written in the form

$$\frac{d^2x}{dt^2} + F(t)x = 0, \tag{8.6}$$

where

$$F(t) = \tfrac{1}{3}(R_{ij}V^iV^j + 2\sigma^2).$$

Then the problem of finding a point q conjugate to S_0 along $\gamma(t)$ becomes that of finding a solution $x(t)$ to eqn (8.6) which vanishes at q (Tipler, 1976). Specifically, for the orthogonal timelike geodesic $\gamma(t)$, a point q along $\gamma(t)$ will be conjugate to S_0 provided a solution $x(t)$ of eqn (8.6) exists which satisfies the initial conditions

$$x(0) = \alpha, \qquad \frac{dx}{dt}\Big|_{t=0} = \alpha{\chi_i}^i, \tag{8.7}$$

and which vanishes at q. In order to analyse the occurrence of zeros in the solutions to eqn (8.6), one can use the Sturm comparison theorem for the solutions of second-order differential equations (see, for example, Hille, 1969) which compares the distribution of zeros of the solutions $u(t)$ and $v(t)$ of the equations

$$\begin{aligned} \frac{d^2u}{dt^2} + G_1(t)u &= 0 \\ \frac{d^2v}{dt^2} + G_2(t)v &= 0, \end{aligned} \tag{8.8}$$

where $G_1 \le G_2$ in an interval (a, b). The Sturm theorem then shows that if $u(t)$ has m zeros in $a < t < b$, then $v(t)$ has at least m zeros in the same interval and the ith zero of $v(t)$ must be earlier than the ith zero of $u(t)$.

Now let $k^2 = \min F(t) = \min \frac{1}{3}(R_{ij}V^iV^j + 2\sigma^2)$ and consider the equation

$$\frac{d^2x}{dt^2} + k^2x = 0. \tag{8.9}$$

Then applying the Sturm theorem to eqns (8.6) and (8.9) we see that if the solution to eqn (8.9) satisfying the initial conditions (8.7) has a zero in the interval $0 < t < t_1$, then the solution of eqn (8.6) defined by the same initial conditions must have a zero in the same interval, which must occur before the zero of the solution of eqn (8.9). The general solution of (8.9) can be written as

$$x = A\sin(B + kt).$$

Let us choose the initial conditions as

$$x(0) = \frac{1}{([{\chi_i}^i]^2 + k^2)^{1/2}}, \qquad \frac{dx}{dt}\Big|_{t=0} = \frac{{\chi_i}^i}{([{\chi_i}^i]^2 + k^2)^{1/2}}, \tag{8.10}$$

where $\chi_i{}^i$ is negative valued on S_0 since the universe is expanding everywhere. It may be possible to envisage scenarios in which the universe might be expanding at some places and contracting in some other regions on S_0; however, we shall not consider such possibilities here. It is then easy to see that the corresponding solution of eqn (8.9) is given as

$$x = \frac{1}{k}\sin(\theta - kt), \tag{8.11}$$

with

$$\theta = \sin^{-1}\left\{\frac{k}{([\chi_i{}^i]^2 + k^2)^{1/2}}\right\}.$$

Thus we have $0 < \theta < \pi/2$ and a zero for x must occur within the interval $0 < kt \leq \pi/2$. Then, using the comparison theorem, we see that the solution of eqn (8.6), as defined by the initial conditions above, must vanish within an interval of time $0 < t < \pi/2k$, i.e. if $\gamma(t)$ is any past directed timelike geodesic orthogonal to S_0, then there must be a point q on $\gamma(t)$, conjugate to S_0, within the above interval.

It is now possible to investigate in general the past extension of arbitrary timelike trajectories from the present epoch S_0 following the standard results discussed in Chapters 4 and 5 (Joshi, 1980, 1986). Let p be an event on S_0 and γ be a past directed, endless timelike curve from $p = \gamma(0)$. Suppose γ can be extended to arbitrary values of proper time in the past, then choose $q = \gamma(\pi/2k)$ to be an event on this trajectory. Then, as discussed in Chapter 4, there exists a timelike geodesic γ' from q orthogonal to S_0 along which the proper time lengths of all non-spacelike curves from q to S_0 are maximized and further, γ' does not contain any conjugate point to S_0 between q and S_0. However, as shown above, any timelike geodesic $\gamma(t)$ must contain a point conjugate to S_0 within the proper time length $\pi/2k$. Since this is not possible, we conclude that no timelike curve from S_0 can be extended into the past beyond the proper time length $\pi/2k$.

The assumption in the above, that $R_{ij}V^iV^j \geq k > 0$ on the hypersurface S_0 and also in the past, may be considered somewhat strong but can be justified on the basis that since the universe is expanding, the density in the past must be higher, or at least non-decreasing. It is, however, possible to weaken this to allow k to decrease in the past and still similar conclusions can be drawn implying upper limits on the past extensions. The idea would be to integrate the quantity $R_{ij}V^iV^j$ in the past and then to use the corresponding theorems on the occurrence of conjugate points. We do not go into those details here as our main purpose is to show that the global techniques do imply some useful upper limits in cosmology in a model independent manner, without recourse to the exact symmetries and assumptions of homogeneity and isotropy. The scenario under consideration may be considered sufficiently general for that purpose.

The above results can be employed to obtain general upper bounds to the age of a globally hyperbolic universe in the following manner. Taking the stress-energy tensor of the perfect fluid form

$$T_{ij} = (\rho + p)u_i u_j + p g_{ij},$$

and using the Einstein equations, we get

$$\begin{aligned} R_{ij}V^iV^j &= 8\pi G(T_{ij}V^iV^j + \tfrac{1}{2}T) \\ &= 8\pi G(\rho + p)(V^4)^2 - 4\pi G\rho + 4\pi G p. \end{aligned}$$

This implies

$$R_{ij}V^iV^j \geq 4\pi G(\rho + 3p). \tag{8.12}$$

If we neglect the pressure p for the cosmological fluid, then we get

$$k^2 = \min \tfrac{1}{3}(R_{ij}V^iV^j + 2\sigma^2) \geq \tfrac{4\pi G\rho}{3}.$$

Then, the maximal possible extension for any timelike worldline from the present epoch into the past, or the maximal possible age of the universe t_{max}, is given by

$$t_{max} = \frac{\pi}{2}\left(\frac{3}{4\pi G\rho}\right)^{1/2} = \pi\left(\frac{3}{16\pi G\rho}\right)^{1/2}, \tag{8.13}$$

within the framework of general globally hyperbolic space-times (see Fig. 56). In the case of radiation-dominated models we can take $p = \rho/3$ and we have

$$t_{max} = \pi\left(\frac{3}{32\pi G\rho}\right)^{1/2}. \tag{8.14}$$

The relationships (8.13) and (8.14) provide upper limits to the age even when allowing for departures from homogeneity and isotropy. It should be noted that the ρ occurring in eqn (8.13) is a global minimum taken over the epoch of constant time for the density which is no longer constant, but is a function of the space coordinates. The average mass density as indicated by the visible galaxies is about 10^{-30} gm cm^{-3}. However, as such, the entire subject of observational determination of the mass-energy density of universe is under active investigation and as indicated above, it is believed that about ten times the visible mass may be in some invisible form in the intergalactic or intragalactic medium. The X-ray observations strongly favour the existence of a hot ionized intergalactic gas within the cluster of

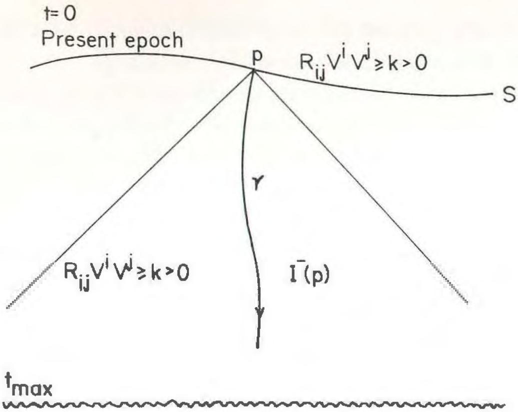

Fig. 56 No timelike curve γ from the surface S extends beyond the maximum limit t_{max} in the past and must encounter a space-time singularity before this epoch.

galaxies, whereas weakly interacting massive neutrinos or other elementary particles such as axions or gravitinos could be another source. We may have to wait for further observations which might determine the contributions from these sources. However, if a hierarchical clustering of matter stops at some stage, say at the scale of the largest structures already observed, then the above density minimum would be achieved. In any case, apart from the matter inhomogeneities, if one accepts the microwave background radiation (MBR) as homogeneously spread out in view of its observed isotropy to a high degree of approximation, and having some kind of global origin, then ρ_{mbr} provides a firm lower limit to the ρ_{min} sought for. Then, the firm general upper limit to the age of the universe as given by eqn (8.14) is

$$t_{max} = \pi \left(\frac{3}{32\pi G\rho_{mbr}} \right)^{1/2} = 3.2 \times 10^{12}\text{yr}, \tag{8.15}$$

where $\rho_{mbr} = 4.4 \times 10^{-34}$ gm cm^{-3}.

Next, if we take the contribution by matter into account, we have to choose an entire range of densities as suggested by the above-mentioned possibilities. The average matter density arising from all possible sources is believed to be anywhere between 10^{-30} and 10^{-28} gm cm^{-3}. In that case, the general upper limits given by eqn (8.13) are 9.43×10^{10} yr and 0.94×10^{10} yr respectively. It is seen that the general upper bound on the age of the universe varies from about 9.4 million years for the highest possible present density, to about 94 million years for the lowest possible value of the density. These limits would appear to be sufficiently tight so as to be useful for many applications in cosmology.

For the Friedmann models, given values of H_0 and q_0 provide the age of universe from eqn (8.1). A comparison of values of ages for the Hubble constant values $H_0 = 40$ km s^{-1}Mpc^{-1} and $H_0 = 120$ km s^{-1} Mpc^{-1}, and $q_0 < 10$, shows that the general upper limits given above are quite tight and interesting in spite of their generality and while allowing for fluctuations from the exact homogeneity. On the other hand, given the lower limits on ages from observations, this suggests several further applications, such as deducing limits on the maximum amount of dark matter density that could be accommodated in the universe, or limits for particle masses such as neutrinos or axions in a model independent way. It is relevant to note in this connection that non-vanishing anisotropy in the universe only contributes towards reducing the age upper limits given above. We discuss some of these applications in the next section.

8.3 Particle mass upper limits

The considerations developed in the previous section can be applied to derive mass upper limits for the elementary particles which were produced in the early universe. We assume here the following scenario. In the very early universe, when the temperatures were arbitrarily high, the particles such as neutrinos or axions and gravitinos were produced abundantly by the particle physics processes such as those mentioned earlier. Subsequently, as the universe cooled, these particles with a small rest mass formed an inert background just as the microwave radiation, and expanded homogeneously together with the radiation. On the other hand, the matter decoupled from the radiation and this neutrino (axions or gravitino) background. This matter then went on to form structures in the universe in the form of galaxies, clusters of galaxies and other higher order structures as pointed out above, making the universe inhomogeneous on the whole even at the scales of hundreds of megaparsecs.

Thus, as implied by the high degree of observed isotropy of the microwave background, we take the radiation to be expanding homogeneously from the early universe. Similarly, once the very high temperature early universe phase is over, the particles such as neutrinos no longer interact with the ordinary matter and expand homogeneously with the radiation. The present number density for these particles in the universe is determined from the present number density of the microwave photons from the statistical arguments. However, accepting the observational evidence that the ordinary matter is spread inhomogeneously at very large scales, we do not assume the universe to be homogeneous and isotropic as a whole, as assumed by the Friedmann solutions. We model such a universe by a general globally hyperbolic space-time (of which the Friedmann model is a special case) which allows for inhomogeneities in the matter distribution on each spacelike surface of cosmic simultaneity. Further, as in the previous section, we assume there is a minimum value for the density on such a surface, which is provided in the present case by the radiation and neutrino backgrounds.

Given this general scenario, the considerations in the previous section and eqn (8.13) allow us to derive upper limits on particle masses in terms of a single parameter, namely, the observed ages of objects in the universe. Given a lower limit on the age of the universe by means of observations on ages of the oldest objects in the universe, such as the globular clusters ages, eqn (8.13) places an upper limit on the ρ_{min}, which includes both the contributions by the microwave radiation and the neutrino background. The knowledge of the present number density of such particles, which is derived from the photon number density, then leads to upper limits on such elementary particle masses.

Observationally, the best lower limits for the age of the universe come from studies of globular clusters of stars in our galaxy. Estimates for this have been given (see, for example, Sandage and Tammann, 1984) to be in the range of 14–20 Gyr (1 Gyr = 10^9 yr.) Symbalisty, Yang and Schramm (1980) have suggested a consistent age estimate in the range 13.8–24 Gyr. Nuclear cosmochronometric studies by Thielmann, Metzinger and Klapdor (1983) suggest a range 18.8–24.8 Gyr. In more recent studies, this range has been somewhat narrowed down. For example, the ages of oldest stars are estimated to be about $13\text{–}17\times10^9$ yr (Sandage and Cacciari, 1990), and the age of elements is taken to be in the range of $12\text{–}16\times10^9$ yr (Cowan, Thielmann and Truran, 1987). In general, it appears reasonable to choose the range of observed ages between 10 and 20 gigayears from nuclear cosmochronometry or such similar direct astrophysical measurements. To get an idea of the numbers involved, if we choose the lower limits on the age of universe to be 10 and 20 gigayears as two extremes, the limits on mass density given by eqn (8.13) are $\rho_{max} = 8.9 \times 10^{-29}$ gm cm^{-3} and $\rho_{max} = 2.2 \times 10^{-29}$ gm cm^{-3} respectively. It is thus seen that despite being sufficiently general so as to allow for inhomogeneities in the universe, these density constraints are fairly tight and interesting.

Based on this consideration, we now derive general mass limits for the neutrinos, gravitinos and the axion, assuming each time that these particles make up the dark matter in the universe. Of course, if the dark matter is composed of more than one component, such as neutrinos and gravitinos, these limits will be clearly tightened. It is also clear that the procedure can be applied to other forms of particles as well and in general it provides the constraints on the dark matter content in the universe. In fact, a remarkable consensus in astrophysics and cosmology today is that non-luminous or dark matter surrounds bright stars and galaxies, and makes up the dominant material content of our universe (Primack, 1984). Evidence for this derives from the (Newtonian) gravitational effect that such matter seems to exert on visible stellar systems. Furthermore, dark matter seems to be present on all distance scales—from the local neighbourhood of our sun and the Milky Way, to clusters and super clusters of galaxies and the expansion scale of the universe itself. For the local solar neighbourhood, explanation of dark matter can be had in terms of brown dwarfs (Jupiter-like objects). However, on larger scales no satisfactory explanation exists in terms of conventional, ordinary matter (namely, baryons and leptons). The search for an answer to this has covered a wide span of possibilities. At the present time, the most likely candidates for dark matter seem to be (a) low mass, faint stars; (b) massive black holes; and (c) massive neutrinos, axions or particles predicted by supersymmetry (Blumenthal, Faber, Primack, and Rees; 1984). Requirement of consistency with the observed abundance of primordial helium discounts possibilities (a) and (b), implying that possibility (c) merits a serious study, although, in fairness, it must be stated that the actual existence of these particles is yet to be established. Such particles are produced in the early universe, and with a small non-zero rest mass, they would be gravitationally dominant today because of their significant number density. Hence, they acquire cosmological significance in relation to the dark matter problem.

Consider first a neutrino dominated universe, where the number density of each species is given as $\sim 110\ \mathrm{cm}^{-3}$. Then ρ_{max} can be worked out using eqn (8.13) corresponding to various observed ages, which provides the upper limit for the neutrino density. Corresponding to the ages 10 and 20 billion years, the limit on sum of neutrino masses over all types satisfies the limits $\sum m_i \leq 453\mathrm{eV}$ and $\sum m_i \leq 112\mathrm{eV}$ respectively (Joshi and Chitre, 1981b). As proposed by Symbalisty *et al.* if higher observed ages up to 24 billion years are allowed then these limits still get much more tighter. It is clear from the laboratory bounds on neutrino masses stated earlier that the above limits are rather interesting even when they are obtained within a general framework allowing for inhomogeneities in the matter distribution in the universe.

Now, assuming that the dark matter is mainly composed of gravitinos, which were produced copiously in the early universe as a result of breakdown of local supersymmetry, we derive mass bounds for these particles (Roy, Joshi and Chitre, 1991). We show here a stable gravitino to respect an upper mass bound ~ 1 keV and an unstable one a lower mass bound ~ 10 TeV from the global hyperbolicity property alone without the assumptions of a Friedmann universe. The specific Friedmann cosmology with a critical mass-density in fact reduces the upper bound for a stable gravitino to ~ 0.1 keV as we shall discuss in Section 8.5.

Cosmological bounds on the gravitino mass have received much attention in the literature (see, for example, Pagels and Primack, 1982; Weinberg, 1982; Khlopov and Linde, 1984; Cline and Raby, 1991; Berezinsky, 1991) and there is a broad consensus in excluding the mass region ~ 1 KeV to ~ 10 TeV. However, the discussion is usually given within the framework of a Friedmann universe. Though the detection of microwave background radiation has made the big bang origin of the universe indispensable, the observation of voids and filaments beyond supercluster scales as pointed out above can make one question the stringent assumptions of the maximally symmetric model. Suppose one wants to relax the assumptions of that model. How are the mass constraints on the gravitino affected? This is the question we address here. Let us again accept the general features of global hyperbolicity; for that is, that the universe admits a foliation by spacelike hypersurfaces at all post-Planckian epochs; however, without assuming that the spacelike hypersurfaces have spherical symmetry or exact homogeneity and isotropy. Now consider a general globally hyperbolic universe uniformly filled with gravitinos of non-zero rest mass $m_{3/2}$ contributing a mass-density $\rho_{3/2}$. Presumably, these gravitinos were copiously produced during the hot early phase of the universe through reactions such as gluon + gluon $\rightarrow$ gluino + gravitino, electron + photon $\rightarrow$ selectron + gravitino, and so on. The gravitinos, if stable, may turn out to be the dominant source of present-day dark matter. In case they are unstable, we shall consider an epoch in which most of them have not decayed yet. Let us now re-examine the arguments leading to the gravitino mass constraints. Our supposing, namely that the dominant contribution to the mass-density of the universe in the epoch under consideration comes from a free non-relativistic gas of gravitinos of mass-density $\rho_{3/2}$, then leads to the inequality

$$R_{ij}V^iV^j \geq \frac{k^2}{2}\rho_{3/2}. \tag{8.16}$$

The above, when combined with eqn (8.13) implies

$$t_{max} = \frac{\pi}{2k}\sqrt{\frac{6}{\rho_{3/2}}}. \tag{8.17}$$

If the gravitino is stable (for example, by way of being the lightest superparticle in an R-parity conserving broken supersymmetric theory), using $t_{age} \leq t_{max}$ where t_{age} is the present age of the universe, wè can write

$$\rho_{3/2} = m_{3/2}n_{3/2} \leq \frac{3\pi M_p^2}{16}\frac{1}{t_{age}^2}, \tag{8.18}$$

where M_p is the Planck mass $M_p = G^{-1/2}$ and $n_{3/2}$ is the present gravitino number density. Moreover, the gravitino number density can be written, following Pagels and Primack (1982), as

$$n_{3/2} = \frac{3}{2} \cdot \frac{43}{11} \frac{1}{g_{3/2d}} n_\gamma . \tag{8.19}$$

In the above, n_γ is the present number density of photons which is ~ 400 cm^{-3} and $g_{3/2d}$ stands for the number of thermalized relativistic degrees of freedom at the gravitino 'freeze-out' temperature. This can be reasonably estimated (Pagels and Primack, 1982) to be ~ 200. The use of this fact and the substitution of eqn (3.19) into eqn (3.18) leads to an upper bound on the stable gravitino mass which varies from ~ 4 keV to ~ 1 keV as t_{age} is varied from 10×10^9 to 20×10^9 years. This bound is quite comparable to that of Pagels and Primack (1982) except that the authors there had used a maximal value of the mass-density as the critical density $\rho_{max} \sim 2 \times 10^{-29}$ gm cm^{-3}, an input which can only be derived from estimates of the Hubble constant H_0 and the deceleration parameter q_0 within a Friedmann universe (see, for example, Weinberg, 1972). Our result, on the other hand, follows from the global hyperbolicity property of the space-time alone. Needless to say, if there are other stable neutral particles in the universe (such as massive neutrinos), these limits will be tightened further.

The situation is quite different for an unstable gravitino. It is subject to decay processes such as gravitino $\to$ gluon + gluino and gravitino $\to$ photon + photino. As the early universe expands and cools, gravitinos begin to decay effectively (that is, without their decay products recombining efficiently) only when the temperature T (times the Boltzmann constant) falls below the gravitino mass. Consider an epoch when the age of the universe is less than but comparable to the gravitino lifetime τ. At this stage the universe can still be taken to be largely gravitino-dominated. Now the global hyperbolicty requirement eqn (8.13), coupled with the information in (8.18), can be restated in terms of the gravitino number density $n_{3/2}(\tau)$ just before substantial numbers of gravitinos decay as the inequality

$$\tau \leq \frac{M_p}{4} \sqrt{\frac{3\pi}{m_{3/2} n_{3/2}(\tau)}} . \tag{8.20}$$

Weinberg (1982) has estimated τ to be $\alpha M_p^2 m_{3/2}^{-3}$, where α is a numerical constant of order unity which depends on the details of the favoured decay process. Thus eqn (8.20) becomes

$$\frac{\alpha^2 M_p^2}{m_{3/2}^5} \leq \frac{3\pi}{16 n_{3/2}(\tau)} . \tag{8.21}$$

The number density of gravitinos $n_{3/2}$ is defined at temperature T. But after they decay, the thermalization of the decay energy changes the temperature to T' where $(kT')^4 = \beta n_{3/2}(\tau) m_{3/2}$, β being an unknown parameter of $O(1)$ involving the number of species of particles with mass below kT, below kT', and below kT_f, where T_f is the gravitino 'freeze-out' temperature. The Weinberg (1982) argument, that the corresponding rise in entropy density must not be allowed to adversely affect the neutron to proton ratio (by producing too much ^{4}He, say), is now valid and requires

$$kT' = \beta^{1/4}[n_{3/2}(\tau)m_{3/2}]^{1/4} > 0.4 \text{ MeV}. \tag{8.22}$$

Combining eqn (8.21) and eqn (8.22) leads to the lower bound

$$m_{3/2} > \alpha^{1/3}\beta^{-1/6}\left(\frac{16}{3\pi}\right)^{1/6}(0.4\text{MeV})^{2/3}M_p^{1/3} \sim 10\,\text{TeV}. \tag{8.23}$$

Thus the Weinberg lower mass bound on the unstable gravitino obtains just from global hyperbolicity and constraints of nucleosynthesis without reference to the Hubble constant or to a Friedmann model of the universe.

We now consider the possibility that the axions constitute the dark matter in the universe. Existence of axions was originally invoked (Peccei and Quinn 1977) in order to explain the property of charge-parity conservation of strong interactions. The axion rest mass m_a is not exactly specified in these theories; it depends on a parameter which has a wide range of possible values. In this connection, useful information can be derived from astrophysical considerations, for example, the requirement that the axion density be less than the critical density for the closure of the universe implies on the basis of the standard Friedmann model that $m_a > 10^{-5}$ eV (Preskill, Wise and Wilczek, 1983; Abbott and Sikivie, 1983; Dine and Fischler 1983). We confine our discussion here to showing that even within the framework of Friedmann models, only one parameter, namely the observed ages of the universe, is enough to get important cosmological limits for axions, rather than requiring the two parameters H_0 and q_0 (Datta and Joshi, 1986).

Now, within the Friedmann model, the contribution to the density due to axions (produced in the early universe) as a function of temperature T is (Preskill, Wise and Wilczek 1983; Abbot and Sikivie, 1983; Dine and Fischler, 1983)

$$\rho_a(T) = \frac{3m_a T^3 f_a^2}{M_p \Lambda_{QCD}}, \tag{8.24}$$

where $\Lambda_{QCD} \approx 200$ MeV is the scale parameter in quantum chromodynamics. The axion mass m_a is related to the vacuum expectation value f_a

of the scalar field that spontaneously breaks the Peccei–Quinn symmetry invoked to explain the CP-invariance of strong interactions, and is given by (Weinberg, 1978; Wilczek, 1978)

$$m_a = 1.24 \times 10^{-5}\text{eV}\left(\frac{10^{12}\,\text{GeV}}{f_a}\right). \tag{8.25}$$

The above equation is independent of cosmology. Particle physics, however, does not specify the exact value of f_a; it can lie anywhere between the weak interaction scale and the mass scale of grand unification.

If dark matter is made up entirely of axions, then using the general arguments above we can write $\rho_a(T) \leq 3\pi/16GT_{age}^2$. This gives, using eqns (8.24) and (8.25),

$$f_a \leq \frac{\pi}{16G}\frac{M_p\Lambda_{QCD}}{T^3}\frac{1}{t_{age}^2}\frac{1}{(1.24\times 10^{-2}\,\text{GeV}^2)} = (f_a)_{max}. \tag{8.26}$$

Taking $T = 2.73$ K as the present temperature of the universe, eqn (8.26) gives a bound based on cosmological as well as particle physics considerations. This can be substituted in eqn (8.25), which is independent of cosmology, to obtain the following lower bound on the axion mass:

$$m_a \geq 1.24 \times 10^{-5}\,\text{eV}\left(\frac{10^{12}\,\text{GeV}}{(f_a)_{max}}\right). \tag{8.27}$$

We find that corresponding to the range of observed ages from (13 to 25) Gyr, the range of upper limits on f_a is (1.15–0.30$\times 10^{12}$ GeV, and the lower limit on m_a is (1.07–4.09)$\times 10^{-5}$ eV. In contrast, the usual arguments based on Friedmann models generally assume that the axion energy density does not exceed the critical density for the closure of the universe, and the knowledge of H_0. The interesting feature of our consideration is a relationship between lower limits on the axion mass and t_{age} without invoking the Hubble parameter.

We have modelled here the cosmological universe by retaining only very general features of the Friedmann models. A general anisotropic and inhomogeneous universe is considered in terms of a globally hyperbolic space-time. Now, the Friedmann model rests on the premises that the universe is isotropic and homogeneous. Although this leads to a picture whose validity must be eventually justified by observations, there is no fundamental physical justification that isotropy and homogeneity (and a spherically symmetric expansion scheme) are strictly obeyed in all the regions of space and at all epochs of time. It appears that the universe is not homogeneous at least up to the scales of 100 Mpc (although isotropy has good observational support). Hence, studies of relativistic cosmological models not obeying homogeneity and isotropy assumptions continue to be important (see, for example, the review by MacCallum, 1979; and Raychaudhuri, 1979). This apart, any conclusion inferred within the framework of Friedmann models have normally required the knowledge of the Hubble parameter H_0 and the deceleration parameter q_0. Keeping in mind the present uncertainties that prevail in the cosmological observations to ascertain the values of these quantities (as well as the inherent restrictive nature of the basic assumptions of the Friedmann models), it would be both interesting and desirable to have some model-independent conclusions regarding the constituents of dark matter as this is a major problem in cosmology today. We have demonstrated here that, if particles such as neutrinos, gravitinos, or axions make up the dark matter in the universe, interesting limits can be set on their properties on the basis of a consideration of space-times in a rather general manner, without recourse to the specific assumptions of isotropy and homogeneity, characteristic of the Friedmann model of the universe. Such limits are derived in terms of the observed ages of universe, which is the only parameter involved in such an approach.

8.4 The cosmological constant

We now derive bounds on the cosmological constant within the framework of a general space-time. The consequences of a non-zero value for cosmological constant in the context of dynamics of the universe have been extensively discussed. In fact, a universal constant Λ can be inserted in Einstein's equations without destroying the general covariance of the theory to write

$$R_{ij} - \tfrac{1}{2}Rg_{ij} + \Lambda g_{ij} = 8\pi G T_{ij}. \tag{8.28}$$

In recent years, there has been a tendency to regard the cosmological constant as one of the most fundamental physical entities (see,for example, Kolb and Turner, 1990). It is generally accepted that the cosmological constant should be considered as part of the stress-energy tensor representing the non-zero vacuum expectation value of T_{ij} generated by quantum fluctuations. Thus, Λ is a number (with unit cm^{-2}) which can, in principle, be calculated from local quantized fields.

On the other hand, the observational methods to determine Λ invoke the departures from the exact Hubble flow for the distant galaxies. The bounds thereby determined for Λ are rather tight, giving $|\ \Lambda\ | \leq 10^{-55}$ cm^{-2} (see, for example, Zeldovich and Novikov, 1983). We note, however, that the above method is beset with several observational uncertainties and stringent assumptions regarding the nature of the universe as a whole (whereas we can only observe part of it), namely, its exact homogeneity and isotropy and an exact measurement of the Hubble constant, which is rather uncertain at the moment. It would, therefore, be highly desirable to have an alternative approach to this problem where the above assumptions can be relaxed. It is the purpose of this section to present such an alternative method to place limits on Λ where we assume neither the global homogeneity and isotropy of the universe nor the values of the Hubble constant H_0. As earlier, the parameter used is the observed ages of the universe.

Using the Einstein equations with a non-zero Λ we compute $F(t)$ in eqn (8.6) to obtain

$$R_{ij}V^iV^j > 4\pi G\rho - \Lambda c^2. \tag{8.29}$$

This gives, for the maximum possible age of the universe t_{max} in the past,

$$t_{max} = \frac{\pi}{2}\left(\frac{3}{4\pi G\rho - c^2\Lambda}\right)^{1/2}. \tag{8.30}$$

This analysis allows for the departures from spherical symmetry, or perturbations from exact homogeneity and isotropy.

We note that observations on the departure from the Hubble law for distant galaxies mentioned earlier only imply that $\mid \Lambda \mid$ is very small. However, Λ could have a positive or negative sign. We consider here each case separately.

For obtaining limits when Λ is negative, we can ignore the contribution from the matter term in eqn (8.30) because including the same would, in any case, tighten the bounds given here. Then eqn (8.30) can be written as

$$t_{max} < \frac{\pi}{2}\left(\frac{3}{c^2 \mid \Lambda \mid}\right)^{1/2}. \tag{8.31}$$

As pointed out earlier, a variety of considerations independent of the cosmological models, such as the ages of stars and globular cluster ages, place the lower limit to the age of universe in the range $(8–24)\times 10^9$ years. Then, $t_{age} < t_{max}$, and eqn (8.31) gives

$$\mid \Lambda \mid = \left(\frac{3\pi^2}{4t_{max}^2 c^2}\right) < \left(\frac{3\pi^2}{4t_{age}^2 c^2}\right). \tag{8.32}$$

Thus, for example, with $t_{age} = 20 \times 10^9$ years, we have,

$$\mid \Lambda \mid_{max} = 2.1 \times 10^{-56}\,\mathrm{cm}^{-2}. \tag{8.33}$$

We note that for the entire range of ages $(10–24)\times 10^9$ years, $\mid \Lambda \mid_{max}$ turns out to be of the order of $10^{-56}\mathrm{cm}^{-2}$. Hence, the limits computed here within a general framework turn out to be actually better than those obtained using the observational method. In fact, this is not unexpected in view of the large uncertainties prevailing today in the measurements to detect the departures from the exact Hubble flow of galaxies H_0.

When Λ is positive, it contributes negatively to the overall stress-energy density in the universe and corresponds to a negative value of pressure. However, if we demand that the overall stress-energy of the universe is positive and the strong energy condition is satisfied, then eqn (8.29) implies

$$\Lambda < \frac{4\pi G\rho}{c^2}. \tag{8.34}$$

With the input of energy densities, the above provides tight upper limits on the cosmological constant values. For example, for $\rho = 5 \times 10^{-29}$ gm cm^{-3}, we get $\Lambda < 4.7 \times 10^{-56}\text{cm}^{-3}$.

As shown in the previous section, cosmological considerations have been very successful in placing stringent limits on the masses of elementary particles, notably those of neutrinos, gravitinos and so on. It was, however, pointed out by Barrow (1981) that most of these results rest on the unverified cosmological assumption of $\Lambda = 0$. It is possible that a non-zero value of Λ will emerge as a result of phase transitions in the early universe. In such a situation, the particle mass upper limits will be revised considerably (Barrow, 1981) as argued below. The assumptions of complete homogeneity and isotropy give the Friedmann equation

$$\frac{\dot{R}^2}{R^2} = \frac{8\pi G\rho}{3} - \frac{k}{R^2} + \Lambda. \tag{8.35}$$

With the usual definitions of the Hubble constant H_0, the deceleration parameter q_0 and dimensionless density parameter $\Omega_0 = 8\pi G\rho_0/H_0^2$ we can write

$$\Omega_0 = \frac{2\Lambda}{3H_0^2} + 2q_0. \tag{8.36}$$

Hence, if $\Lambda > 0$ the matter density can be considerably higher without adversely affecting the value of q_0 and particle masses can be correspondingly larger.

Now, the observational data on the lower limits for ages of stars, globular clusters, the red shift magnitude, diagram and the quasar data imply that when $\Lambda \neq 0$, the parameters q_0 and Ω_0 obey the following limits

$$-4.4 < q_0 < 5.6, \qquad 0.05 < \Omega_0 < 0.94.$$

In such a situation, we have an upper limit on ρ_0

$$\rho_0 < 1.7 \times 10^{-28} h_0^2 \text{gmcm}^{-3},$$

where the parameter h_0 reflects the uncertainty in the measurement of the Hubble constant H_0 and is taken to be $1/2 < h_0 < 1$. This weakens the particle mass limits considerably. Taking the case of light neutrinos, we have

$$m_\nu < 280 h_0^2 \mathrm{eV},$$

instead of the value $m_\nu < 60 h_0^2\,\mathrm{eV}$ for the $\Lambda = 0$ case. The bounds for heavy neutrinos, heavy leptons and so on, will be analogously revised.

We can similarly use the global consideration in a general scenario to obtain particle mass limits when the cosmological constant is non-zero. It can be seen that the particle mass limits obtained in the previous section are revised upwards, depending on the value of Λ in the case of a positive cosmological constant. Thus, it turns out that a non-zero cosmological constant significantly changes the elementary particle mass constraints in cosmology. Clearly, the case for a non-zero value of the cosmological constant is still open.

The Friedmann model for the universe is beset with serious problems, such as the existence of horizons which separate the universe into causally disjoint regions. How would such regions achieve the required exact homogeneity is an unsolved problem. The inflationary scenario (Guth, 1981; Barrow and Turner, 1982; Linde, 1991), postulating an exponential expansion in the early universe, was proposed to solve these problems of the standard cosmological model. In this picture, a scalar field with a non-zero vacuum energy density serves as a driving force for the exponential expansion. This can be interpreted as a large positive cosmological constant in the early universe, which would reduce to its present small positive value after the phase transition in the early universe. We now derive general limits on such a positive cosmological constant (or equivalently the vacuum energy density) within the framework of an inflationary universe, again using the general assumption of global hyperbolicity but without the use of any special properties like spherical symmetry or homogeneity of the underlying space-time. Such a picture may be useful when the observed inhomogeneities are to be taken into account by the inflationary models. A clear upper limit of 1/3 is obtained for the vacuum energy density parameter Ω_v, while the lower limit is found to depend on the age of the oldest object in the universe.

The inflationary scenario, in its several available versions, claims to resolve the cosmological conundrums such as the horizon problem, monopole problem and at the same time predicts that the density parameter Ω_0 of the universe must be close to unity. All these models require a fine tuning of the net cosmological constant, during and after the inflationary phase.

This cosmological constant is equivalent to the vacuum energy density ρ_v,

$$\rho_v = \frac{\Lambda c^2}{8\pi G}. \tag{8.37}$$

One could appeal to observations on departures from the exact Hubble law for distant galaxies and related cosmological phenomena to get more information or constraints on this remnant vacuum energy density. However, it is also possible to determine the range of values for ρ_v that may exist today without assumptions of exact symmetries or crucial dependence on the present value of H_0 (Joshi, Padmanabhan and Chitre, 1987). In other words, we permit departures and small perturbations from the maximally symmetric model while assuming the global hyperbolicity as earlier. In particular, the flat ($k = 0$) Friedmann models with small perturbations satisfy this criterion.

Einstein's equations with the cosmological constant can be rewritten as

$$R^i{}_k - \tfrac{1}{2}\delta^i{}_k R = 8\pi G (T^i{}_k)^{eff}, \tag{8.38}$$

where

$$(T^i{}_k)^{eff} = T^i{}_k - \frac{\Lambda c^2}{8\pi G}\delta^i{}_k. \tag{8.39}$$

Using the above equations and eqn (8.13), and the procedure outlined earlier, we obtain for the maximum possible age,

$$t_{max} = \frac{1}{4}\sqrt{\frac{\pi}{G}}\sqrt{\frac{3}{\rho_m - 2\rho_v}}. \tag{8.40}$$

We emphasize that only global hyperbolicity was assumed in obtaining the above result. We further assume that the cosmological constant is a residue from an inflationary phase. Since inflation implies $\rho = \rho_c$, where ρ_c is the critical density, we can write, $\rho_c = \rho_m + \rho_v$. For a dust-filled universe, the pressure term arises purely from vacuum energy density, $p = -\rho_v$. Defining

$$\Omega_v = \frac{\rho_v}{\rho_c}, \qquad \Omega_m = \frac{\rho_m}{\rho_c},$$

we get

$$t^2_{max} = \frac{3\pi}{16G}\frac{1}{\rho_c - 3\rho_v}, \tag{8.41}$$

or, in other words,

$$1 - 3\Omega_v = \frac{3\pi}{16G\rho_c t^2_{max}} = \frac{1}{2}\left(\frac{\pi}{H_0 t_{max}}\right)^2. \tag{8.42}$$

Since the right-hand side in the above is positive, a clear upper limit on Ω_v is given if we assume that the positive cosmological constant is not so large as to dominate the positive matter density by negative pressures and that the strong energy condition is respected,

$$\Omega_v < \tfrac{1}{3}. \tag{8.43}$$

Further, $t_{age} < t_{max}$ where t_{age} is the age of the oldest known objects in the universe (that is, globular clusters). Combining these constraints with eqn (8.43) we get the bounds

$$\tfrac{1}{3}[1 - \tfrac{1}{2}(\pi/H_0 t_{age})^2] < \Omega_v < \tfrac{1}{3}. \tag{8.44}$$

Scaling $t_{age} = 1.5p \times 10^{10}$ years and $H_0 = 100h_0\ \mathrm{km\ s^{-1}\ Mpc^{-1}}$, the above equation can be written as

$$\tfrac{1}{3}(1 - 2/h_0^2 p^2) < \Omega_v < \tfrac{1}{3}. \tag{8.45}$$

It is seen that a positive lower bound on Ω_v is obtained for $ph_0 > \sqrt{2} \approx 1.4$. Thus, for example, for $h_0 = 1$, observations of ages as high as 21 billion years would imply a positive lower bound on Ω_v. As the consistent age limits for the universe are usually proposed to be in the range 13.8–24 billion years, it is conceivable that Ω_v is bounded from below. On the other hand, if quasars are taken to be the oldest objects in the universe, having ages of the order of 30 billion years, then eqn (8.45) would provide definitive lower limits for Ω_v.

The scenario assumes a new dimension if dynamical considerations are introduced. It is well known that $\rho = \rho_c$ cannot be achieved by luminous matter alone. The dark matter which has to be invoked can come in many forms, conveniently classified as 'cold' or 'hot'. If either one of these choices prove to be successful, still our conclusions would be valid; however, investigations show that neither 'cold' nor 'hot' matter by itself can account for observations at all scales. This has motivated the study of scenarios in which a cold dark matter candidate decays in the recent past. In such models the present-day universe will be radiation-dominated because of the existence of relativistic decay products.

In such a situation, we must have

$$p = \tfrac{1}{3}\rho_m - \rho_v,$$

giving, in place of eqn (8.41),

$$t_{max}^2 = \frac{3\pi}{16G}\,\frac{1}{2\rho_m - 5\rho_v}. \tag{8.46}$$

This, in turn, leads to the following bounds on Ω_v:

$$\frac{2}{7}[1 - (\pi/2H_0 t_{age})^2] < \Omega_v < \tfrac{1}{3}. \tag{8.47}$$

Again following the earlier convention, it is seen that the above gives a positive lower bound to Ω_v when $ph_0 > 1$. It is clear that for a range of values of p and h_0 this inequality will be satisfied and hence Ω_v would be bounded from below.

We have derived here bounds on the cosmological constant, or on the vacuum energy density in an inflationary scenario using general properties of a space-time and without invoking the exact symmetries of the underlying model. In particular, we do not demand the spacelike hypersurfaces evolving in time to have any special properties like spherical symmetry or homogeneity and isotropy, except to satisfy the criterion of global hyperbolicity. It is then possible to derive the required bounds in terms of the ages of the oldest objects in the universe.

8.5 Friedmann models revisited

We have shown here in the earlier sections that interesting limits on the particle masses and other properties can be obtained within a general framework without invoking the exact symmetries, and using only one parameter, namely the observed ages of oldest objects in the universe. This has the advantage of allowing for the perturbations and inhomogeneities in the form of structures in the universe. On the other hand, the usual considerations on this issue are limited to Friedmann models and use two observed parameters, namely the Hubble constant H_0 and the deceleration parameter q_0; or H_0 and ρ_0 and so on. Now, Friedmann models are a special case of general globally hyperbolic space-times. It is thus natural to ask if the conclusions in the Friedmann models could be obtained using just one parameter, namely the observed ages, and if so whether that improves the earlier conclusions. We discuss this question here and show that in fact this is possible.

In Friedmann universes, the present age t_{age} of the universe is given by eqn (8.1) in terms of the parameters H_0 and q_0. Using eqns (8.2) and (8.3) the age of the universe can be expressed as

$$t_{age} = \left(\frac{3}{4\pi G\rho_0}\right)^{1/2} \sqrt{q_0} f(q_0). \tag{8.48}$$

If we now consider a situation at the present matter-dominated epoch where we can specify the mass density of the universe and treat q_0 as

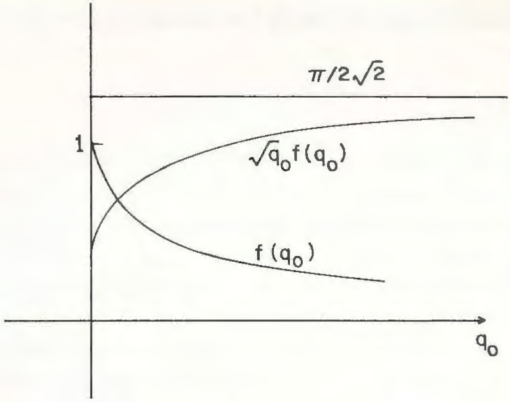

Fig. 57 The behaviour of the function $\sqrt{q_0}f(q_0)$.

a parameter, then the behaviour of the function $\sqrt{q_0}f(q_0)$ shows it as a monotonically increasing function of q_0, which approaches a maximum limit of $\pi/2\sqrt{2}$ as q_0 tends to infinity (Fig. 57). This can be easily seen using the expression for $f(q_0)$ given above. Thus, the maximum possible age t_{max} of a closed Friedmann universe ($q_0 > 1/2$) is given by

$$t_{max}^{closed} = \frac{\pi}{2\sqrt{2}}\left(\frac{3}{4\pi G\rho_0}\right)^{1/2} = \pi\left(\frac{3}{32\pi G\rho_0}\right)^{1/2}. \tag{8.49}$$

In the same manner the maximum possible age for open models ($0 < q_0 < 1/2$) is given by

$$t_{max}^{open} = \frac{\sqrt{2}}{3}\left(\frac{3}{4\pi G\rho_0}\right)^{1/2} = \left(\frac{1}{6\pi G\rho_0}\right)^{1/2}. \tag{8.50}$$

If we choose the observed range of ages as 13.8 to 24 billion years, these estimates on the age allow us to use eqns (8.49) and (8.50) to obtain the required particle mass limits. For example, if we consider the neutrino again in three types, namely the electron, muon, and tau neutrinos and anti-neutrinos, with the number density for each type given by $\sim 110\ \mathrm{cm}^{-3}$, the upper bounds can be obtained from the requirement that $t_{age} < t_{max}$. This requirement implies, using eqns (8.49) and (8.50), that the maximum allowed mass-energy density for both closed and open models is given by (Joshi and Chitre, 1981a)

$$\rho_{max}^{closed} = \frac{3\pi}{32Gt_{age}^2}, \qquad \rho_{max}^{open} = \frac{1}{6\pi Gt_{age}^2}. \tag{8.51}$$

For the closed models, we can deduce from the above relations that for the age of 10 billion years, the bound on neutrino masses is given by $\sum m_i < 227\,\mathrm{eV}$, and for the age of 20 billion years this bound is rather tight, given by $\sum m_i < 56$ eV. For other ages the bounds are corresponding scaled. It is clear that these bounds are tighter than the bounds usually obtained in the Friedmann models using the values of H_0 and q_0. In fact, if the ages such as 24 to 25 billion years are allowed, it follows that a very small value for the neutrino rest mass is allowed. In the case of an inflationary model with the mass density given by $\rho = \rho_c$, and $q_0 = 1/2$, the above numbers again provide the mass bounds, taken as the limiting values of the closed models.

For the open models of the universe, the corresponding mass limits are much tighter. For the age of 10 billion years, the bound on $\sum m_i$ is $\sim 40\,\mathrm{eV}$ where as the bound corresponding to the age 20 billion years is $\sim 10\,\mathrm{eV}$. It is readily seen that the data on the age of the universe places fairly tight limits on neutrino masses. Our results indicate that experimental findings of neutrino masses of the order of a few tens of electron-volts appear inconsistent with the lower limits on the age as implied by nucleo-cosmochronology. The upper limits on the neutrino masses in the open universe and the data on the age of the universe thus suggest the possibility that the presence of neutrinos with a mass of even a few electron-volts could dominate the mass of the universe leading to is closure.

Clearly, the non-zero rest mass of neutrinos, gravitinos, or axions produced in the early universe has several important implications for some of the outstanding astrophysical problems such as the solar neutrinos, the missing mass in clusters of galaxies, and the closure of the universe. We must await the experimental limits on the neutrino mass to be firmly established before we can draw any further conclusions about the role of neutrinos in influencing the structure and evolution of the universe at large.

For the case of gravitinos considered in the previous section, if one does specifically assume a Friedmann universe with a critical value of the mass density, a certain amount of tightening of the bounds becomes possible. Now eqn (8.17) gets replaced by

$$t_{age} = \frac{1}{H_0} f\left(\frac{k^2 \rho_{3/2}}{6H_0^2}\right),$$

where f is the function of the deceleration parameter q_0 which has been explicitly written out in the argument of f. If the mass-density of the gravitino-dominated universe is critical, $\rho_{3/2} = 3k^{-2}H_0^2$ and since $f(1/2) = 2/3$, one gets

$$t_{age} = \frac{1}{k}\left(\frac{4}{3\rho_{3/2}}\right)^{1/2}. \tag{8.52}$$

The product $\rho_{3/2}.t_{age}^2$ is now a factor $9\pi^2/8 \approx 11.1$ smaller than the upper bound on it following from eqn (8.18). Hence, the corresponding upper bound on the mass of a stable gravitino gets reduced by the same factor that is, there is an order of magnitude tightening to $\sim 0.1\,\mathrm{keV}$. A similar argument can be given for the lower mass bound on an unstable gravitino. Here there is an enhancement but through the factor $(9\pi^2/8)^{1/6}$ which is only ~ 1.5. Thus, the gravitino mass exclusion region now extends from $\sim 0.1\,\mathrm{keV}$ to $\sim 10\,\mathrm{TeV}$.

It may be noted that we have assumed here the absence of any significant inflation-induced dilution of the gravitino abundance (see, for example, Khlopov and Linde, 1984). The latter can be used to evade the bounds obtained here, but a heavy price has to be paid for the same, namely a low reheating temperature destroying our present understanding of baryogenesis, as concluded by Khlopov and Linde.

To summarize, we have shown that cosmological bounds on the particle masses do not need to invoke the (maximally symmetric) Friedmann universe but are rather generally valid in any globally hyperbolic space-time. Thus, in case the existence of very large scale inhomogeneities leads to the abandonment of the exact symmetries of the Friedmann model, the mass upper limits for particles such as neutrinos, axion, and gravitinos still stand in a reasonably tight form. On the other hand, if one is willing to assume a Friedmann universe and a critical mass-density, a certain tightening of the bounds is possible as has been shown here.

9

QUANTUM EFFECTS NEAR THE SPACE-TIME SINGULARITY

Any discussion on space-time singularities cannot be considered complete without consideration of quantum effects in the vicinity of the same. The reason is, when the curvatures grow arbitrarily high near a space-time singularity, especially when it is a scalar polynomial or strong curvature singularity, the classical general relativity can be relied upon less and less and the quantum effects must become important and may dominate. In fact, the very existence of singularity could become questionable, when quantum effects are taken into account. For example, DeWitt (1967) suggested the possibility for singularity avoidance within a quantum frame-work by imposing the condition $\psi[{}^3g] = 0$ at the singular boundary, where ${}^3g$ denotes a three-geometry. Thus, the state functional for the universe vanishes at the singularity.

Unfortunately, despite serious attempts over the past half century, no quantum theory of gravity is available presently which one could use to examine the quantum effects near a classical singularity, or to study the question of whether the singularity can be totally avoided in a full quantum gravity theory. As a result, many different approaches have been tried out to study this problem, such as quantizing the homogeneous and isotropic Friedmann models, or quantizing only part of the degrees of freedom of the metric tensor (see, for example, Gotay and Demaret, 1983). The basic difficulty here is, in the absence of a consistent and complete quantum gravity theory there is no unique scheme to apply. As was anticipated by Brill (1975), there are a variety of ways of viewing the singularity and the quantum theory of universe, and depending upon the procedure used, the singularity may or may not be avoided. Hence, we must try out various alternatives as well as possible model calculations which may eventually lead to a better insight on the status of singularities in quantum gravity.

It is in this spirit that we consider in this chapter a somewhat limited approach to study the quantum effects near a space-time singularity, where one quantized only a limited range of fluctuations of the metric tensor, namely the conformal degree of freedom of g_{ij}. This approach, initiated by Narlikar (1981, 1984) was further developed by Narlikar and Padmanabhan (1983), Joshi and Narlikar (1986), and Joshi and Joshi (1987, 1988). Using both the path integral and operator methods one could see here that the quantum effects must diverge in the vicinity of a classical space-time singularity, offering the possibility that the singularity could be avoided when the quantum effects are taken into account. Of course, it remains an open question as to how the results will be altered when the quantization of other degrees of freedom is taken into account and due to this factor this approach may not be considered complete in its content. However, it has the merit of deducing definite conclusions and to that extent it provides some insight into the workings of quantum effects near a space-time singularity. Another basic purpose for us to undertake this discussion here is to point out the applications in this area of certain global concepts and methods which we have discussed in the earlier chapters, and to show how various general features can be deduced. It would seem that even when the quantization of other degrees of freedom is included, several basic global features of space-time will continue to be relevant, especially in a semi-classical approximation where one considers the propagation of quantum matter fields on a fixed classical back ground. In that sense, the approach given here, namely the quantization of the conformal factor, could also be viewed as the quantization of a scalar field on a fixed space-time background.

In Section 9.1 we discuss basic issues and global aspects in quantum gravity. The quantization of conformal degree of freedom is described in Section 9.2. The evolution of these quantum effects near the space-time singularity is then worked out in Section 9.3 using the path integral method in a general globally hyperbolic space-time and using the fact that at the classical level the conformal fluctuations of metric leave the causal structure of space-time invariant. Despite the elegance of path integral approach, it is beset with problems such as the question of measure on the space of all paths. On the other hand, the operator approach in quantum theory is more direct and has an intuitive appeal. We address the question of evolution of quantum effects using the operator approach in Section 9.4 and a wide range of space-times are covered to examine the quantum effects near the singularity. Section 9.5 considers the question of measure of singular geometries in the space of all allowed geometries when the quantum effects are taken into account in the above manner. The above discussion so far has been in the cosmological context. In Section 9.6, the final section, we examine the quantum effects near a singularity forming in the homogeneous gravitational collapse, namely near the Schwarzschild singularity at $r = 0$ within a black hole.

9.1 Basic issues in quantum gravity

The major issue in generating a quantum theory of gravity is that all the familiar methods such as the covariant or canonical quantization, which have worked very well in quantizing other physical fields, fail completely and run into serious problems when applied to the gravitational field. As we will observe, many of these difficulties are related to the fact that global issues and features become extremely important as soon as we try to describe the gravitational field in detail. As a result, novel ideas and techniques are required and at the same time there are several profound conceptual issues as well which need a serious attention.

Why should one quantize the gravitational field and have to make a transition from the classical general relativity to a quantum theory of gravitation? Several compelling arguments can be given in this connection. Of the four fundamental forces of nature known to us, namely the electromagnetic field, the weak and strong nuclear forces, and gravitation, the first three have been quantized. Hence, if one desires a unified description of all the forces of nature, then gravity also must be quantized. The idea that all the fundamental forces and interactions in nature are described by a single, unified conceptual framework and are manifestations of a single entity is extremely appealing. Such a unification for the electric and magnetic forces was achieved by Maxwell in the last century in his theory of electromagnetic fields. Then the Weinberg–Salam theory has provided a unified quantum treatment for the electromagnetic and weak interactions, called the electroweak theory. It is generally believed now that the grand unified theories also include the strong nuclear forces in this general scheme. Thus, the aesthetic motivation to include the gravitational field as well in this grand scheme is quite compelling. In fact, Einstein himself spent considerable effort towards unifying his theory of gravitation with the electromagnetic theory.

Both the general relativity and quantum theory have been extremely successful in their respective domains, namely the macroscopic gravitating systems and the microscopic world of atoms and elementary particles. Both have been able to explain the observed phenomena and have made predictions which have been verified subsequently to a great degree of accuracy. However, in their conceptual framework they are completely different from each other without having any similarities and without giving any clues to the possible common underlying structures. General relativity assumes the arena of a space-time as the basic framework in which all the physics takes place and assumes a continuum differentiable manifold structure for the same in which the space-time metric defines the causal structure and the light cones determines the propagation for the particles. The ten independent components of this symmetric metric tensor g_{ij} are supposed to obey the Einstein equations, which is a system of partial differential equations involving the second derivatives of the metric tensor. In fact, the components of the metric tensor are to be obtained by solving the Einstein equations which construct the space-time. On the other hand, the quantum theory is based on a Hilbert space of the state vectors, which describe any given state of the quantum mechanical system. Physical fields are represented by linear operators which act on these vectors to produce real numbers which are interpreted as the results of measurements.

The real explanation for this diversity in approach could lie in the fact that as compared to other forces mentioned above, the force of gravity is extremely weak. The strength of gravitational force between two elementary particles is much weaker as compared to other forces between them. The scale at which the classical description breaks down for a theory depends on the masses and charges of the particles involved and the values of the fundamental constants involved. Thus, for the quantum theory of gravity, the length scale at which quantum effects will become important will be determined by the values of the velocity of light c, the Planck constant $\hbar$, and the Newtonian constant of gravity G. A unique length scale is set by these three constants which is given by $\ell_p = (G\hbar/c^3)^{1/2}$, which is called the *Planck length*. In cgs units it has the value of the order of 10^{-33} cm which is the fundamental length scale at which the quantum gravity effects will be important. It is a coincidence that the values of the fundamental constants are such that the Planck length ℓ_p is so small and the corresponding Planck energy is so high as compared to the laboratory scales at which we normally operate. This allows us to use the general relativity for the purposes of astrophysics and cosmology and the quantum theory for atomic and subatomic physics. Even though they look so different at the laboratory scale, the experience so far in physics tells us that both must be different approximations of the same quantum gravity theory when suitable limits are taken. This unified quantum gravity theory will become important at the Planck length scales and Planck energies.

There are two reasons why such a quantum gravity theory could have observational implications in spite of the very high energies involved. The first reason is that the Planck energies are only a few orders of magnitudes higher than the energies involved in the grand unified theories, and thus the possibility is not ruled out that quantum gravity could have useful implications for the grand unified theories of weak and strong interactions. The second reason is provided by the general relativity theory itself, which predicts the existence of space-time singularities where the curvatures and densities could be unboundedly high as we have pointed out in Chapters 5 and 6. It follows that near the space-time singularity and near the Planck length scale the quantum gravity will be important and the physics very near the singularity will be described by the quantum gravity only. Thus, one needs quantum gravity to describe the very early universe near the big bang singularity, which provided the origin of the universe. The behaviour of physical fields in this phase will have observational implications for the later evolution of the universe and the observations subsequently made.

Several approaches have been tried out to quantize the general theory of relativity, such as the covariant quantization, the canonical method, the path integral approach and such others (we refer to Isham, Penrose and Sciama (1981), Hawking and Israel (1979), and Ashtekar and Stachel (1991) for reviews on various methods involved and for a detailed discussion on basic problems in quantum gravity). Attempts have also been made to modify the general relativity theory, such as the higher-order gravity. None of these ideas have worked successfully and they run into one or other difficulty which turns out to be formidable. The major reason behind all such difficulties appears to be the very unconventional structure of general relativity as compared to other fundamental theories of physics. In general relativity the metric tensor plays a dual role. The components g_{ij} determine the kinematic arena in that the metric determines the causality, light cone structure and spacelike hypersurfaces of the space-time. On the other hand, it is a dynamical quantity which has a role equivalent to the Newtonian potential as well. Thus, there is no a priori background given in general relativity in which one propagates the physical fields. For example, for the Maxwell theory the background is taken as the Minkowski space-time in which the electromagnetic fields are characterized by the field tensor F_{ij}. The flat space-time provides a kinematic framework in which the flat metric η_{ij} defines the light cones and spacelike surfaces in which the field propagates. Given the values of the field on one spacelike surface, the values on a future spacelike surface can be determined. In contrast to this, in general relativity no back ground space-time is given but the space-time itself has to be constructed by solving the Einstein equations for the metric potentials. These metric potentials then determine the space-time as a kinematic arena and play the role of dynamical variables also. Such a dual role basically is a characterization of the equivalence principle and gives the theory its elegance, but it also raises its own conceptual difficulties when one tries to quantize gravitational force.

The covariant approach tries to solve this quantization problem by mimicking the method for electromagnetic fields stated above. Here the space-time metric of general relativity is broken into two parts, $g_{ij} = \eta_{ij} + \gamma_{ij}$, where η_{ij} is the flat metric and the γ_{ij} part contains rest of the non-linearities of the metric tensor. There is a great conceptual simplicity in this method in that to first order in γ_{ij} this gives just a free spin-2 field theory in the Minkowski space-time. Including the full non-linear γ_{ij} reduces the general relativity to a self-interacting spin-2 field theory in the flat Minkowski space-time. Now one could use the familiar perturbation methods of quantum field theories to obtain a formal perturbation series.

The major trouble that arises here is that the perturbation theory one generates in this manner for the γ_{ij} is not renormalizable. The squared Planck length ℓ_p^2 enters the perturbation series as an expansion parameter and in order to achieve finiteness, it is required that the series be finite in each order. The difficulties in achieving this are pointed out by Deser, van Nieuwenhuizen and Boulware (1975).

The more recent supergravity theories represent an initiative in a similar spirit. Here the idea is to couple the gravity with suitably chosen matter fields. The hope is then the infinities of the bosonic fields, including those of gravity, will be cancelled by those of the fermionic fields and one would get a renormalized theory of gravity interacting with matter. The most refined theory in this effort, called the $N = 8$ supergravity, has indeed several good features such as some cancellation of infinities does actually occur, the Hamiltonian is manifestly positive and the theory is unitary as well. The drawback, however, is that again the theory is not renormalizable (see, for example, Hawking and Rosek, 1981, for a review on these developments).

The canonical quantization scheme can be used for a theory which can be cast in the Hamiltonian form. Now, this is possible for the general theory of relativity (see, for example, Wald, 1984, for details of this formalism). One describes the state of the system by wave functions Ψ of the configuration variables and the time evolution of the system is determined by the Schrödinger equation which uses the Hamiltonian operator derived from the classical Hamiltonian of the theory. The serious stumbling block here has been a constraint equation which must be solved in order to reduce to the variables which represent only the true dynamical degrees of freedom. Several attempts have been made to get rid of this constraint problem in canonical gravity (see, for example, the review by Kuchar, 1981, for details). In this connection, a new initiative has been provided recently by Ashtekar (1986) where a set of new variables is introduced to deal with the equations of canonical quantum gravity. In the usual set of canonical variables (q_{ij}, p^{ij}), the canonical equations are non-polynomial and cannot be solved in the full quantum gravity. With the introduction of the new variables, all such equations become polynomials and this offers a range of possibilities to address the traditional questions which could not be solved earlier. We refer to Ashtekar (1991) and references therein for recent developments on this topic. Particularly interesting here has been the loop space representation developed by Smolin and co-workers, in which case the quantum states can be expressed as functionals of closed loops on a spatial three-dimensional manifold. This allowed them to generate an infinite-dimensional space of solutions for all the quantum constraints which includes the Fock space of states in the weak field limit.

Similarly, the path integral approach has been tried out to quantize gravity, where the major problem turns out to be the definition of measure on the space of all the paths in consideration. One could foliate the space-time here and consider on each spacelike surface all possible three-geometries h_{ij}. The amplitude in question here is that of going from a given spatial metric on the initial hypersurface to another given spatial geometry at the final spacelike hypersurface. Each of such amplitude is represented by the action functional and the path integral is the sum over all such possible paths. However, the foliation given for the space-time is not unique but is quite arbitrary and so the initial and final times given here are rather arbitrary. Thus, the physical meaning of the amplitude here is not clear apart from the question of measure stated above.

Apart from the above methods, several other approaches have also been tried out to achieve the quantization of gravitation fields. These include the Euclidian path integral approach of Hawking, and the twistor approach initiated by Penrose. It has also been tried to modify Einstein equations so that the resulting quantum gravity turns out to be renormalizable via the covariant approach. Here the Lagrangian is modified by adding terms quadratic in curvature which gives the field equations involving the fourth derivatives of the metric tensor. As it turns out, these approaches bring with them their own difficulties of one kind or an other.

An interesting recent effort towards quantum gravity has been the development of string theories in particle physics (Green, Schwarz and Witten, 1987). Here, instead of point particles and associated fields, one considers one-dimensional objects, called strings, as fundamental physical objects. Then familiar physical particles, including the zero rest mass spin-2 gravitons arise as different components of the string excitations. This theory has only one free parameter and hence it really needs a very minimal external input with the coupling fixed automatically. As a result, this theory could have been the theory of everything, containing all physical information. Actually, the string model turns out to be unitary and there is a general agreement that it is finite order by order in the perturbation theory. The trouble, however, is when summed, the perturbation series diverges quite powerfully. Of course, such divergences are also there in theories such as the quantum electrodynamics. But quantum electrodynamics already assumes the flat Minkowski background and continuum space-time structure and so on, and one could blame such infininities on similar external factors. On the other hand, the string theories leave very little outside and must solve the problem of infinities within the theory itself.

The present state of affairs and the history of efforts in quantum gravity indicate that the usual perturbative approach that has worked so well in particle physics and quantum field theories does not appear to yield desired results towards quantization of the gravitational force. It appears almost certain that before we could formulate a working theory of quantum gravity, profound conceptual issues and problems involving global and topological concepts must be addressed, which concern the very fundamental nature and structure of the space-time itself. It certainly appears necessary to understand in a better manner the non-perturbative aspects of gravity as implied by the basic approach in general relativity, and the role that they might play in quantum gravity.

What are the global aspects which are likely to be very important while formulating and studying a quantum theory of gravity? Probably, it is our assumption of the smooth continuum structure for the space-time that is causing most of the problems encountered so far. What is the structure of space-time at microscopic levels and at the scales near the Planck length? The considerations on space-time topology could also play an important role. Is it reasonable to work with a fixed space-time topology in quantum gravity? Will quantum fluctuations not cause fluctuations in space-time topology as well? For a discussion on these issues and possible relationship between space-time topology and quantum gravity, we refer to Isham (1991) and Friedman (1991) and references therein. Other important problems are those of the issue of time in quantum gravity and formulating the quantum theory of the universe as a whole (Hartle, 1986).

To summarize, the emphasis today in the attempts towards quantization of the gravitational force certainly involves non-perturbative and global questions concerning the structure of space-time itself in quantum gravity. Without such an effort progress in this field does not appear to be possible. In fact, it is possible that in the string theory itself non-perturbative methods might lead to a solution of the present problems. For a recent survey on conceptual problems in quantum gravity, we refer to Ashtekar and Stachel (1991).

It must be emphasized, however, that in spite of the major issue of the quantum gravity remaining unsolved, the theory of a quantized free matter field in a fixed curved background is a well-defined problem which has been studied in detail. Though this is only an approximation to the full quantum gravity with quantum matter fields, it gives an idea of possible quantum effects in full quantum gravity. In particular, this theory predicts the creation of particles by the gravitational field, and when applied to the case of a black hole geometry this particle creation implies a thermal spectrum of emission with a temperature proportional to the surface gravity of the black hole (Hawking, 1975). This effect shows an interesting relationship between black holes and thermodynamics. For a review of these developments we refer to Wald (1984).

9.2 Quantizing the conformal factor

Let us denote by (M, g) the space-time manifold whose geometry satisfies the classical Einstein's equations. The energy tensor T_{ij} is taken as that for a system of particles minimally coupled to gravity—an assumption that may very well be justified in the very early universe if all the other basic interactions are less important than gravity in view of asymptotic freedom.

The manifold with metric conformal to g will be denoted by $(\overline{M}, \Omega^2 g)$ or simply by M_ϕ, where

$$\phi = \Omega - 1 \tag{9.1}$$

is the conformal fluctuation from the classical space-time metric. Under the transformation $g_{ij} \to \Omega^2 g_{ij}$ we get

$$R \to R(1+\phi)^{-2} + 6(1+\phi)^{-3}\Box\phi. \tag{9.2}$$

The classical Hilbert action accordingly transforms to

$$S_H = \frac{1}{16\pi}\int_V R\sqrt{-g}\,d^4x \to \frac{1}{16\pi}\int_V [(1+\phi)^2 R - 6\phi_i\phi^i]\sqrt{-g}\,d^4x, \tag{9.3}$$

where V is the space-time four-volume over which the action is defined, x^i are the four coordinates, and we have set $c = 1, G = 1, h = 1$. Here ϕ_i denotes the gradient $\partial\phi/\partial x^i$, and the indices are raised or lowered by the classical metric g.

Normally such a conformal transformation should lead to the existence of second derivatives of ϕ arising from the $\Box\phi$ term in eqn (9.2). However, Green's theorem can be used to transform them to the form given in eqn (9.3) together with a surface term defined over the boundary of V. This surface term cancels a similar term which comes from the conformal transform of the surface term introduced by Gibbons and Hawking (1977) to effectively remove the second derivatives of the metric tensor from the Hilbert action. Henceforth we shall ignore these surface effects.

To the Hilbert action must be added the term describing a system of non-interacting particles $a, b, \ldots$ of masses $m_a, m_b, \ldots$

$$S_m = \sum_a \int m_a ds_a \tag{9.4}$$

where s_a is the proper time of a-th particle. It is clear that

$$S_m \to \sum_a \int m_a \Omega ds_a = S_m + \sum_a \int m_a \phi ds_a \tag{9.5}$$

under the conformal transformation.

Writing $S = S_H + S_m$ for the total classical action, we therefore find that under the conformal transformation, S transforms to

$$S \to S_\phi = S + \frac{1}{16\pi}\int_V (R\phi^2 - 6\phi_i\phi^i)\sqrt{-g}\,d^4x. \tag{9.6}$$

This action functional of ϕ plays the key role in conformal quantization. Notice that there are no linear terms in ϕ because the classical action satisfies the stationarity condition $\delta S = 0$. To proceed further we show how the classical geometrodynamics are modified by the quantum fluctuations. Suppose M is foliated by a sequence of space-like hypersurfaces Σ and we define a time coordinate t to label Σ as $t =$ const. Denote by $\{^3G\}$ the three-geometry on Σ. Then according to Isenberg and Wheeler (1979), the solution of Einsteins equations may be considered as a sequence of three-geometries ^3G on Σ sandwiched between two hypersurfaces $t = t_I, t = t_F, t_I < t_F$. The four-volume V is thus the slab of space-time between these initial and final hypersurfaces which we shall label Σ_I and Σ_F respectively. The appropriate initial conditions to be specified on Σ_I are the conformal part of $\{^3G\}$ and the trace of the extrinsic curvature tensor, K.

In abstract terms we may identify any sequence of ^3G from Σ_I to Σ_F as a 'path' Γ in function space. The classical path Γ_c denotes a particular sequence that satisfies the Einstein equations and the given initial conditions. Any other path may not satisfy Einstein's equations and hence will not be an acceptable description of classical gravity.

This situation is altered in the following way in conformally quantized gravity. Let $\Omega^2 g$ be the new metric in V for an arbitrary conformal factor Ω. Suppose that we allow Ω to vary as a function of $x^i (x^0 \equiv t)$ in $t_I \leq t \leq t_F$ subject to the requirement that $\Omega = \Omega_I(\phi = \phi_I)$ on $t = t_I$ and $\Omega = \Omega_F(\phi = \phi_F)$ on $t = t_F$. Such variation gives a restricted set S of paths Γ, over which we sum the Feynman path amplitude $\exp iS[\Gamma]$ to obtain the propagator

$$K[\phi_F, t_F; \phi_I, t_I] = \sum_{\Gamma \in S} \exp iS[\Gamma]. \tag{9.7}$$

The classical path Γ_c corresponds, of course, to $\phi = 0$ and is not included in S. (We exclude here the trivial situation wherein a constant ϕ does give a uniformly scaled classical solution. Such a case would be included in S only if $\phi_I = \phi_F$.) Given the state of the system $\phi = \phi_I$ at the initial epoch, we could evolve this state into past using the propagator (9.7) to obtain the state of the universe at an earlier time $t = t_F$ (see Fig. 58).

From (9.7) we can at once write down K as a path integral

$$K[\phi_F, t_F; \phi_I, t_I] = \exp iS_c \int \exp\left\{\frac{i}{16\pi}\int_V (R\phi^2 - 6\phi_i\phi^i)\sqrt{-g}\,d^4x\right\} D\phi. \tag{9.8}$$

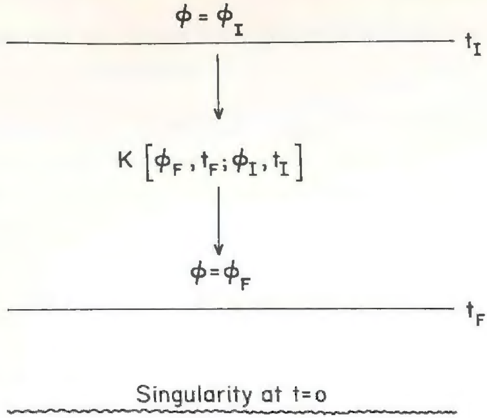

Fig. 58 The quantum evolution of states in the past.

Here S_c is the action evaluated for the classical path Γ_c. The quantities R, g, and so on, are supposed to be known from the classical solution while ϕ is the quantum input. The propagator K tells us how the quantum conformal fluctuations behave in the four-volume V, and contains the entire content of conformal quantization.

The evaluation of eqn (9.8) has been done (see, for example, Narlikar, 1981, 1984). Skipping the details given therein we state the final answer,

$$K[\phi_F, t_F; \phi_I, t_I] = F(t_F, t_I) \exp \frac{3i}{8\pi}\chi, \tag{9.9}$$

where

$$\chi = \sum\sum\nolimits_{P,Q=I,F} \int\int A_{PQ}(\mathbf{r}, t_P; \mathbf{r}', t_Q)\phi_P(\mathbf{r})\phi_{\mathbf{Q}}(\mathbf{r}')\mathbf{d}^3\mathbf{r}\mathbf{d}^3\mathbf{r}'. \tag{9.10}$$

In the above expression $\mathbf{r}$ and $\mathbf{r}'$ stand for the three space-like coordinates. Recall that since ϕ_I and ϕ_F have been specified respectively on the hypersurfaces $t = t_I$ and $t = t_F$, they are functions of three (space-like) coordinates only.

The coefficients A_{II} and A_{FF} are related to the advanced and retarded Green's functions of the operator $\Box + R/6$ in a somewhat implicit form. The cross coefficient $A_{IF} \equiv A_{FI}$ is, however, simpler to state:

$$A_{IF}^{-1} \equiv A_{FI}^{-1} = G^R(F, I). \tag{9.11}$$

where G^R, the retarded Green's function between $(\mathbf{r_F}, \mathbf{t_F})$ and $(\mathbf{r_I}, \mathbf{t_I})$ satisfies the wave equation

$$\Box_x G^R(X, B) + \tfrac{1}{6}R(X)G^R(X, B) = [-\bar{g}(X)]^{-1/2}\delta_4(X, B) \tag{9.12}$$

δ_4 being the four-dimensional delta function.

9.3 Evolution of quantum effects

The propagator K obtained in Section 9.2 can be used in the following way. Let $\Psi_I(\phi_I)$ denote the wave functional characterizing the state of geometry at $t = t_I$ in terms of quantum theory. Then the state at $t = t_F$ is given by

$$\Psi_F(\phi_F) = \int K[\phi_F, t_F; \phi_I, t_I]\Psi_I(\phi_I)D\phi_I. \tag{9.13}$$

Even without calculating this functional integral the following important conclusion can be drawn from it. First we note that if $G^R(X, B)$ diverges as X approaches the classical singular epoch in (M, g), then $A_{IF} \to 0$. Hence the cross term in χ tends to zero and the dependence of $\Psi_F(\phi_F)$ on $\Psi_I(\phi_I)$ is only through a constant factor. In other words the final state 'totally loses memory' of its initial state.

This divergence of quantum uncertainty indicates that the classical solution itself is not a reliable indicator of quantum reality. In the set S of non-classical paths, there is, however, a set of paths with geometries that are conformal to g and also singular. The complement of this set in S is made of paths with geometries that, while conformal to g, are non-singular. In the diverging range of conformal fluctuations which set predominates? We shall discuss this question in Section 9.4.

In the present section, we examine the evolution of quantum effects in the limit of approach to the space-time singularity. A particularly simple illustration is presented by the standard big bang model. Keeping the universe homogeneous and isotropic, the line element is given by the Robertson–Walker form

$$ds^2 = -dt^2 + S^2(t)\left[\frac{dr^2}{1 - kr^2} + r^2(d\theta^2 + \sin^2\theta\, d\phi^2)\right] \tag{9.14}$$

where $k = 0, 1$, or -1 and (r, θ, ϕ) are the co-moving coordinates of a typical fundamental observer. The function $S(t)$ vanishes at the singular epoch $t = 0$.

It is worth noting that the most general fluctuations of eqn (9.14) keeping the homogeneity and isotropy of the universe are the conformal fluctuations with ϕ a function of t only. Then functionals of ϕ and eqn (9.13) are simplified to ordinary integrals:

$$\Psi_F(\phi_F) = \int K[\phi_F, t_F; \phi_I, t_I]\Psi_I(\phi_I)d\phi_I. \tag{9.15}$$

The explicit behaviour of K is given by Narlikar and Padmanabhan (1983). If Ψ_I is given by the wave packet

$$\Psi_I(\phi_I) = (2\pi\triangle_I^2)^{-1/4}\exp(-\phi_I^2/4\triangle_I^2), \tag{9.16}$$

then Ψ_F is given by a similar function of ϕ_F but with $\triangle_I$ replaced by

$$\triangle_F = \triangle_I\left(A_{II}^2 + \frac{1}{16\triangle_I^4}\right)^{1/2} A_{IF}^{-1} \tag{9.17}$$

The divergence of $\triangle_F$ as $t \to 0$ arises from the fact that

$$A_{IF} \sim S_c(t) \quad as \quad t \to 0 \tag{9.18}$$

where $S_c(t)$ is the classical Friedmann expansion factor. Since all singular solutions of the form (9.14) must have $\phi(t)/\triangle_F \to 0$ as $t \to 0$, the zero measure of singular solutions follows.

The above treatment, however, requires the wave function of the universe to be a Gaussian wave packet. In fact, we show now that it is possible to generalize the above result to the case when the state of the universe is no longer assumed to be a Gaussian wave packet, and examine the behaviour of quantum conformal fluctuations near the classical singularity. Though it is reasonable to demand that the classical situation $\phi = 0$ should be an average of the quantum description, the Gaussian packet assumption is obviously too strong a demand for the wave function of the universe. We shall thus choose the state of the universe to be represented by a general wave functional $\Psi(\phi, t)$ which satisfies the normal requirements such as $\int \Psi\bar{\Psi}D\phi = 1$. An analysis of this case may yield some insights for the general situation to be dealt with next.

As pointed out by Narlikar and Padmanabhan (1983), the propagator for the case of homogeneous and isotropic space-times is given as :

$$K(\phi_F, t_F; \phi_I, t_I) = F(t_F, t_I)\exp\frac{3i}{8\pi}[A_{II}\phi_I^2 + 2A_{IF}\phi_I\phi_F + A_{FF}\phi_F^2] \tag{9.19}$$

where we have $A_{IF} \to 0$ as $t \to 0$, that is, in the limit as the classical singularity is approached. Let $\Psi(\phi, t)$ be the general wave function of the universe whose time evolution is given by eqn (9.15). We shall continue to impose the requirement that the classical state $\phi = 0$ should at all times be the average of the quantum description. Therefore we have at $t = t_I$

$$\langle\phi_I\rangle \equiv \int_{-\infty}^{\infty} \bar{\Psi}_I(\phi_I)\phi_I\Psi_I(\phi_I)d\phi_I = 0. \tag{9.20}$$

Further, the requirement $\langle\phi\rangle = 0$ at all times means that the 'momentum' conjugate to ϕ averages to zero. At $t = t_I$ this condition implies that

$$\int_{-\infty}^{\infty} \bar{\Psi}'_I(\phi_I)\Psi_I(\phi_I)d\phi_I = 0 = \int_{-\infty}^{\infty} \bar{\Psi}_I(\phi_I)\Psi'_I(\phi_I)d\phi_I \tag{9.21}$$

where the overhead prime denotes differentiation.

The average $\langle\phi_F\rangle$ at any later time t_F can now be shown to be zero provided eqns (9.20) and (9.21) hold. For, by the relation (9.15) we obtain

$$\begin{aligned}\langle\phi_F\rangle = &\int_{-\infty}^{\infty} \bar{\Psi}_I(\phi_1)\bar{K}(\phi_F, t_F; \phi_I, t_I)\phi_F \\ &\times K(\phi_F, t_F; \phi_I, t_I)\Psi_I(\phi_2)d\phi_1 d\phi_2 d\phi_F.\end{aligned} \tag{9.22}$$

If one uses the δ-function representations

$$\delta(\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{i\omega x}dx, \tag{9.23}$$

$$\delta'(\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} ixe^{i\omega x}dx, \quad \text{etc.}, \tag{9.24}$$

and

$$\int_{-\infty}^{\infty} G(x)\delta'(x)dx = -G'(0), \tag{9.25}$$

then eqn (9.22) can be simplified as

$$\begin{aligned}\langle\phi_F\rangle = &\frac{\pi i}{2A_{IF}^2} \mid F(t_F, t_I) \mid^2 \\ &\times \int_{-\infty}^{\infty} [\bar{\Psi}(\phi, t_I)\Psi'_I(\phi, t_I) + 2iA_{II}\phi\bar{\Psi}(\phi, t_I)\Psi_I(\phi, t_I)]d\phi.\end{aligned} \tag{9.26}$$

Hence, using eqns (9.20) and (9.21) we get

$$\langle\phi_F\rangle = 0, \tag{9.27}$$

that is, at all times the classical state is the average of the quantum ensemble.

Now the dispersion Δ_F at time t_F is given by

$$\begin{aligned}\Delta_F^2 \equiv \langle\phi_F^2\rangle &= \int_{-\infty}^{\infty} \bar{\Psi}_F(\phi_F)\phi_F^2\Psi_F(\phi_F)d\phi_F \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mid F(t_F, t_I) \mid^2 \exp[iA_{II}(\phi_1^2 - \phi_3^2)] \\ &\quad \times \bar{\Psi}_I(\phi_3, t_I)\Psi_I(\phi_1, t_I)\phi_F^2 \exp[2i\phi_F A_{IF}(\phi_1 - \phi_3)]d\phi_1 d\phi_3 d\phi_F.\end{aligned} \tag{9.28}$$

Again, using eqn (9.24) we get

$$\int_{-\infty}^{\infty} \phi_F^2 \exp[2i\phi_F A_{IF}(\phi_1 - \phi_3)]d\phi_F = \frac{2\pi}{8A_I F^3}\delta''(\phi_1 - \phi_3) \tag{9.29}$$

Using the normalization condition

$$\frac{\pi \mid F(t_2, t_1) \mid^2}{A_{IF}} = 1, \tag{9.30}$$

implied by

$$\int_{-\infty}^{\infty} \bar{\Psi}_F(\phi_F)\Psi_F(\phi_F)d\phi_F = 1 \tag{9.31}$$

we finally get for eqn (9.28)

$$\triangle_F^2 = \frac{1}{4A_{IF}^2}\int_{-\infty}^{\infty}[\bar{\Psi}_I\Psi_I'' - 4A_{II}^2\phi^2\bar{\Psi}_I\Psi_I + 2iA_{II}\phi\bar{\Psi}_I\Psi_I']d\phi \tag{9.32}$$

Since the integral itself is independent of t_F we see that as $t_F \to 0$,

$$\triangle_F{}^2 \sim A_{IF}^{-2}.$$

This dependence is the same as eqn (9.17) for wave packets.

The above derivation uses the path integral method to propagate the wave functions. It is also possible to use the Hamiltonian formulation to examine the behaviour of the dispersion $\triangle$ where one can proceed by constructing the Hamiltonian and a differential equation to evolve the dispersion. Again the results are valid for a general wave function and $\triangle_F$ is seen to diverge near the singularity. We will discuss this approach in the next section.

It is well accepted, however, that the above assumptions of homogeneity and isotropy represent very specialized situations and in general in the actual universe small or large deviations from these must be admitted.

Hence we shall now consider the general situation of a globally hyperbolic space-time where one is no longer constrained by such special requirements and inhomogeneities and anisotropies in the mass-energy distribution are allowed for. Here (M, g) may contain many singularities, either local or cosmological. As pointed out earlier, these space-times form a general class with the only topological constraint that M be homeomorphic to $S \times \mathbb{R}$ where S is a three-dimensional spacelike hypersurface. This also provides a foliation of the space-time by means of spacelike hypersurfaces. We note that this general class includes most of the important cosmologies such as the Friedmann–Robertson–Walker types, the Bianchi models and steady state cosmologies as well as exact solutions like the Schwarzschild space-time.

Normally, one would like to think that a singularity of one type, such as, say, timelike geodesic incompleteness, would imply a singularity of other types also such as null incompleteness. However, this does not necessarily happen and one can find examples of space-times which may be timelike geodesically complete but null incomplete or which are non-spacelike geodesically complete but may contain finite length timelike curves without end points. Thus, it would seem that the notion of geodesic incompleteness is not a sufficient condition that one would like to associate with a genuine physical singularity, which is that the space-time curvatures must grow so large that the local laws of physics are drastically altered and even the usual ideas of space and time may break down. One would like to think of the geodesic incompleteness as an 'effect' rather than a 'cause'. Thus, we shall deal here with curvature singularities which are genuine and physically all embracing in the sense that all observers, whether freely falling or accelerated, falling within the singularity, are destroyed by infinite tidal forces and the curvature scalars or the Riemann tensor components should grow unboundedly along the trajectories falling into the singularity. This is characterized by saying that for any singularity in M (which is not a part of the space-time), all the non-spacelike geodesics falling within the singularity are future incomplete at the singularity and all the non-geodesic non-spacelike curves ending there have finite length without a future end point in M.

This notion is best formulated by means of the proper and terminal indecomposable past (future) sets in space-times as defined in Sections 4.6 and 5.1, which can characterize the singularities as well as the points at infinity of the space-time in terms of an additional boundary attached to M. Since the construction of this boundary involves only the causal structure of space-time, no additional assumptions for M are needed.

As stated earlier, any TIP (or TIF) is generated by a future directed (past directed) non-spacelike curve γ which has no future (past) end point. The singularities within the ideal boundary ∂M can now be distinguished in the following manner: Let P be a TIP of ∂M and Γ be the set of non-spacelike curves generating P. If $L(\gamma) < \infty$ for all $\gamma \in \Gamma$, then P is called a *singularity* of M. Thus, if P is a singularity, any $\gamma \in \Gamma$ will have a finite length and no future end point. On the other hand, if there exists at least one timelike curve with infinite length generating the TIP P, then P is a non-singular TIP (see Fig. 38). Now the assumption stated earlier regarding singularities can be formulated as follows:

All the singularities in M must be singular TIPs.

Thus, if γ is an incomplete non-spacelike future directed geodesic, then the TIP $I^-(\gamma)$ defines a singularity S and our condition tells that for any other future inextendible non-spacelike curve $\lambda \subset I^-(\gamma)$, λ has finite length in future, has no future end point, and $I^-(\lambda)$ is also a TIP defining S. Clearly, λ also terminates into the same singularity S in future. One could now further add a condition on curvature growth as defined in Section 5.4 to ensure that this is a strong curvature singularity where curvatures and energy density would blow up.

With this framework in mind, we now consider the conformal transformations of the original globally hyperbolic geometry (M, g). When one considers arbitrary metric fluctuations in the original geometry (M, g), the causal future and past $I^{\pm}(x)$ are changed in M, the non-spacelike geodesics of the new geometry are different and in general the entire causal structure of M is altered. However, as discussed in Chapter 4, if the metric fluctuations are conformal to the original geometry, the causal relationships remain unchanged and the null geodesics are invariant as point sets under $g \to \Omega^2 g$. But the timelike geodesics of (M, g) and $(M, \Omega^2 g)$ could be quite different. Since the sets $I^{\pm}(p)$ are invariant under conformal fluctuations, the point set defined by any non-spacelike geodesic in (M, g) will in general define some non-spacelike curve for $(M, \Omega^2 g)$ which need not be geodetic. As a consequence, the geodesic completeness properties of conformally transformed space-times are greatly altered and these have been studied by various authors in detail.

For null geodesics, the effect of conformal fluctuations is easy to see since they remain point-wise fixed. Using the geodesic equation, it is seen that under $g \to \Omega^2 g$ the affine parameter k along a null geodesic γ transforms as:

$$k' = \int_0^k \Omega(\gamma(k))dk, \tag{9.33}$$

and hence the completeness properties of null geodesics will be changed. Similar effects can be analysed for non-spacelike geodesics generally and Seifert (1971) derived the following result which we shall use here.

Theorem 9.1. Let (M, g) be a globally hyperbolic space-time. Then there exists a conformal factor $\Omega_c > 0$ such that all the non-spacelike geodesics in $(M, \Omega_c^2 g)$ are complete.

Further, Clarke (1971) constructed a conformal factor Ω_c using a weaker assumption of strong causality such that $(M, \Omega_c^2 g)$ will be null geodesically complete (briefly written as null g-complete). A further improvement was given by Beem (1976) by the construction of Ω_c for M satisfying the non-imprisonment condition, which is a causality condition weaker than strong causality, such that all non-spacelike geodesics in the conformally transformed manifold $(M, \Omega_c^2 g)$ will be complete. The non-imprisonment condition means that whenever a future or past inextendible non-spacelike curve λ enters a compact space-time region K, then λ must leave K.

Now let S denote a singularity in (M, g), that is, it is a singular TIP and let some future directed non-spacelike γ be a generator of $S; I^-(\gamma) = I^-(S)$. Let the metric be conformally transformed as $g \rightarrow \Omega_c^2 g$. Then $(M, \Omega_c^2 g)$ is non-spacelike g-complete and since the causality is invariant under conformal transformations, it is globally hyperbolic. We show that as a result, the singularity S of the original space-time no longer remains a singular TIP but is thrown off as a point at infinity.

For that, first note that $I^-(S)$ remains invariant as a point set and as a TIP under the conformal transformation. Let any non-spacelike curve γ be a generator of this TIP. Choose any $p \in \gamma$. Let q_i be other event on γ to the future of p. Then as pointed out in Section 4.4, there exists a maximal non-spacelike geodesic γ_i of $(M, \Omega_c^2 g)$ from p to q. Let $q_i \rightarrow S$ along γ, then by the above construction we obtain a non-spacelike geodesic γ' of $(M, \Omega_c^2 g)$ which is totally contained in $I^-(S)$ and such that $I^-(\gamma) = I^-(\gamma')$; that is, it is a generator of the TIP S which is future endless (see Fig. 59). However, all the inextendible non-spacelike geodesics of $(M, \Omega_c^2 g)$ are complete and hence γ' has infinite length in future as measured by its affine parameter. Thus, there exists a generator γ' for the TIP S which has infinite length and hence S is no longer a singularity, it is a non-singular TIP in $(M, \Omega_c^2 g)$. It is now clear that all the singularities S in the original (M, g) will be similarly transformed as points at infinity, though they are still within the boundary of M, as they are non-singular TIPs or the points at infinity.

As shown above, the non-spacelike curve γ' of $(M, \Omega_c^2 g)$ is contained within $I^-(S)$ and the same situation continues under conformal transformations, though its status as a geodesic may change to an arbitrary non-spacelike curve. Thus, in (M, g), the point set of γ' will define a non-spacelike curve γ'', totally contained in $I^-(S)$ and $I^-(\gamma'')$ defining the singular TIP S. Hence, in (M, g), the curve γ'' must have a finite length in future being the generator of a singular TIP:

$$L(\gamma'') = \int_p^q \mid g(\partial/\partial t, \partial/\partial t) \mid^{1/2} dt < \infty. \tag{9.34}$$

Using the arc length as parameter along γ'' we have

$$g_{ij}(dx^i/ds)(dx^j/ds) = -1,$$

which gives

$$L(\gamma'') = \int_p^q ds < \infty, \tag{9.35}$$

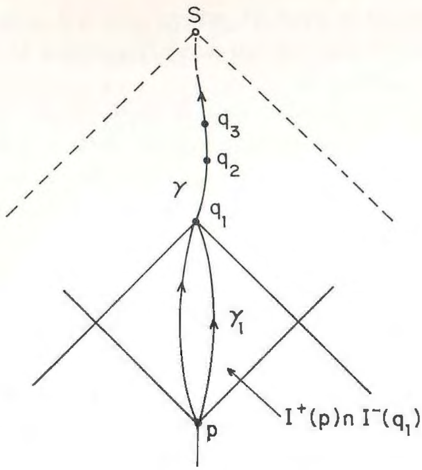

Fig. 59 The event $q_1 \in I^+(p)$ and γ_1 is a non-spacelike geodesic maximizing the lengths of all non-spacelike curves from p to q_1. Taking $q_i \to S$ one can construct a maximal timelike geodesic approaching the singularity S.

in the limit as q tends to the singularity S along γ''. Now, under $g_{ij} \to \Omega_c^2 g_{ij}$, the non-spacelike curve γ'' is mapped into the non-spacelike geodesic γ' of $(M, \Omega_c^2 g)$, which being geodesically complete has infinite length:

$$L(\gamma') = \int_p^q \Omega_c ds = \infty, \tag{9.36}$$

as $q \to S$. Thus, the conformal factor must blow up sufficiently rapidly along γ''. Again, we could have chosen any other generator γ_1 of the TIP $I^-(S)$ and a similar construction could have been repeated, deducing the blow-up of the conformal factor Ω_c along a non-spacelike curve γ_1'' in (M, g) in $I^-(S)$ as the singularity S is approached along γ_1'', using the geodesic completeness of $(M, \Omega_c^2 g)$. We have thus shown that there is an 'overall' divergence in Ω as the singularity S is approached from $I^-(S)$, which has the effect of throwing the singularity off to the infinity.

It may be noted in this connection that the various singularities in M would generally lie at various epochs in the space-time as defined by the foliation of M and that they would be normally covered by event horizons in order that the global hyperbolicity of M is preserved.

Consider now the behaviour of Green's functions in the vicinity of singularities within this general scenario. Since we consider universes with arbitrary distribution of particles and allow inhomogeneities and anisotropies, we no longer have $\phi = \phi(t)$ only. Instead, ϕ becomes a general function of space-time points: $\phi = \phi(x^\mu, t)$, where $\mu = 1, 2, 3$ denotes space coordinates. The evolution of the general wave functional $\Psi[\phi(x^\mu), t]$ in this case is governed by eqn (9.15).

Suppose (M, g) is a general globally hyperbolic space-time satisfying Einstein's equations and with singularities distributed at various epochs as formulated above. Let Ω_c be a conformal factor such that the geometry $(M, \Omega_c^2 g)$ is geodesically complete. Denote by $G(X, B)$ and $G_c(X, B)$ Green's functions for the geometries (M, g) and $(M, \Omega_c^2 g)$ respectively. It is known that Green's functions of two conformally related space-times are connected by

$$G(X, B) = \Omega_c(X)\Omega_c(B)G_c(X, B). \tag{9.37}$$

Let a space-time singularity S in (M, g) be approached along some generator γ in $I^-(S)$. Then, as shown earlier, as $X \to S$ along γ, $\Omega_c(X) \to \infty$ in $I^-(S)$; that is, Ω_c blows up as the singularity is approached. Now in the conformally transformed space-time the singularity S is thrown off to infinity and hence $G_c(X, B)$ will be regular for any X along γ within the space-time. Thus (9.37) implies that Green's function $G(X, B)$ must diverge as Ω_c in the limit of approach to the singularity. From the argument given in the beginning of this section, we therefore conclude that the quantum uncertainty diverges in this limit.

9.4 The operator approach

The path integral method used above is elegant in its approach but has certain problems associated with it such as the issue of measure on the space of all paths (see, for example, Feynman and Hibbs, 1965; Dewitt, 1979). In view of the absence of a full quantum gravity theory, one desires to use methods which have the merit of being transparent and direct. The familiar operator approach in quantum theory satisfies this requirement, which is intuitively much more appealing. In the following, we present some more results examining the quantum effects near a singularity using this approach while quantizing the conformal degree of freedom again. First, we discuss some important exact models before dealing with a space-time of certain generality.

In order to examine how the quantum effects evolve during the final stage of the gravitational collapse in the vicinity of the space-time singularity, we consider first the dust-ball solution as given by Oppenheimer and Snyder (1939) in the form

$$ds^2 = -dt^2 + Q^2(t)\left[\frac{dr^2}{1 - \alpha r^2} + r^2(d\theta^2 + \sin^2\theta\, d\phi^2)\right]. \tag{9.38}$$

In this interior solution, (r, θ, ϕ) are the constant comoving coordinates of a typical dust particle whereas t represents its proper time. The parameter α is related to the starting density ρ_0 of the object and is given by

$$\alpha = 8\pi G\rho_0/3. \tag{9.39}$$

The object is confined to a radius $r \leq r_b$, and its coordinate volume is given by

$$V = \int_0^{r_b} \frac{4\pi r^2 dr}{(1-\alpha r^2)^{1/2}}. \tag{9.40}$$

The collapse is characterized by the scale parameter $Q(t)$ which monotonically decreases from an initial state $Q_0 = 1$ to zero as the singularity is approached, satisfying the differential equation

$$\dot{Q}^2 = \alpha[(1-Q)/Q]. \tag{9.41}$$

We now allow for the conformal fluctuations of this classical system given by

$$g_{ij} \to \Omega^2 g_{ij} \equiv (1+\phi)^2 g_{ij}. \tag{9.42}$$

The function $\Omega > 0$ here is a general function of space-time coordinates; however, in order to be consistent with the exact symmetry of the problem which is homogeneous evolution of dust, we have $\phi = \phi(t)$ only. The idea now is to quantize this system with a conformal degree of freedom. Then, even though classically the state $\phi = 0$ can occur with certainty, quantum mechanically all the other states $\phi \neq 0$ can occur with finite probability. We note that although the metric (9.38) is obtained as an exact solution to the Einstein equations, which corresponds to the $\phi = 0$ situation, the other states in the quantum ensemble need not satisfy the same and they might well be non-singular. The state of the universe now will be represented by a wave function $\psi = \psi(\phi, t)$; however, in the earlier stages of collapse, the situation is highly classical and the states with $\phi \neq 0$ will appear only with vanishing probability. Thus, our task will be to evolve the quantum uncertainty around the classical situation in the limit of approach to the classical singularity. This will be achieved in the following manner. We write down the Lagrangian function as well as the Hamiltonian H for the ensemble above having the conformal degree of freedom. Basically, we are interested in the behaviour of the mean square deviation $\triangle^2$ of ϕ treated as an operator as the collapse progresses.

Defining an operator by

$$A = \phi^2, \qquad \chi = \langle A \rangle, \tag{9.43}$$

we can evolve χ using the time evolution equation (see, for example, Messiah, 1973)

$$i\frac{d\langle A\rangle}{dt} = \langle [A,H]\rangle + i\left\langle \frac{\partial A}{\partial t}\right\rangle. \tag{9.44}$$

It turns out that χ diverges in the limit of approach to the singularity, and hence the state $\phi = 0$ becomes less and less probable as the dispersion around it diverges. Again, consider a system of non-interacting material dust particles $a, b, \ldots$, then the matter action S_m can be written as

$$S_m = -\sum_a m_a ds_a = -\int_{\mathcal{V}} \bar{\rho}\sqrt{-g}\Omega\, d^4x = -\rho_0 V \int_{t_1}^{t_2} (1+\phi)dt. \tag{9.45}$$

Substituting from eqns (9.3) and (9.45) into the total action (9.6) while noting that for the classical solution the scalar curvature is given by $R = 8\pi G\rho_0/Q^3$, we get

$$S = \frac{V\rho_0}{2}\int_{t_1}^{t_2}\left(\phi^2 - \frac{2Q^3}{\alpha}\dot{\phi}^2 - 1\right).dt \tag{9.46}$$

Hence, the Lagrangian function L for the system with the conformal degree of freedom is obtained as

$$L = \frac{V\rho_0}{2}\left[\phi^2 - \frac{2Q^3}{\alpha}\dot{\phi}^2 - 1\right]. \tag{9.47}$$

The associated conjugate momentum is given by

$$p = \partial L/\partial\dot{\phi} = k_1 x^{3/2}\dot{\phi} \tag{9.48}$$

where $k_1 = -2V\rho_0/\alpha$, and $x = Q^2$. The Hamiltonian for the system can now be written as

$$H = \dot{\phi}p - L = (1/2k_1)p^2x^{-3/2} - (k_1\alpha/4)(1-\phi^2). \tag{9.49}$$

To treat the system quantum mechanically, let us now represent the state of the universe by a wave function $\psi = \psi(\phi, t)$. Classically, the state $\phi = 0$ occurs with unit probability. Even though this is no longer the situation now in the sense that there will always be a spread around the classical state, we assume that the classical state, $\phi = 0$, continues to be the average of the quantum ensemble, that is, $\langle\phi\rangle = 0$, while we allow other states to occur with finite probability. The spread in the wave function or the quantum uncertainty is given by

$$\chi \equiv (\Delta\phi)^2 = \langle\phi^2\rangle - \langle\phi\rangle^2. \tag{9.50}$$

The uncertainty in the conjugate momentum is similarly defined and is denoted by ω. As a result of our condition $\langle\phi\rangle = 0$, we write

$$0 = \frac{d}{dt}\langle\phi\rangle = i\int \bar{\psi}[H\phi - \phi H]\psi dt.$$

Using eqn (9.49) as well as the operator representation $p = (1/i)(\partial/\partial\phi)$, it is not difficult to see from the above that

$$0 = \frac{x^{-3/2}}{k_1}\langle p\rangle. \tag{9.51}$$

Thus, $\langle p\rangle = 0$ and we have

$$\chi = \langle\phi^2\rangle, \qquad \omega = \langle p^2\rangle. \tag{9.52}$$

Using eqns (9.43) and (9.44) now, it is possible to evolve χ in time:

$$i\frac{d\chi}{dt} = i\frac{d}{dt}\langle A\rangle = \langle[\phi^2, H]\rangle + i\left\langle\frac{\partial A}{\partial t}\right\rangle. \tag{9.53}$$

The commutators of the operators ϕ and p with the Hamiltonian H are defined in the usual manner and are given by

$$[\phi, H] = \frac{i}{k_1}x^{-3/2}p. \tag{9.54}$$

$$[p, H] = -\frac{i\alpha k_1}{2}\phi \tag{9.55}$$

Substituting in eqn (9.53) together with $\langle\partial A/\partial t\rangle = 0$, we have

$$\frac{d\chi}{dt} = \frac{x^{-3/2}}{k_1}\langle p\phi + \phi p\rangle. \tag{9.56}$$

In order to obtain the differential equation satisfied by χ, we perform one more differentiation by defining

$$\frac{d\chi}{dt} = \langle B\rangle, \qquad B = \frac{x^{-3/2}}{k_1}(p\phi + \phi p). \tag{9.57}$$

Then

$$i\frac{d^2\chi}{dt^2} = i\frac{d}{dt}\langle B\rangle = \langle[B, H]\rangle + i\left\langle\frac{\partial B}{\partial t}\right\rangle. \tag{9.58}$$

Where, using eqns (9.54) and (9.55), we get

$$\langle [B, H] \rangle = -\alpha i x^{-3/2} \langle \phi^2 \rangle + (2ix^{-3}/k_1^2) \langle p^2 \rangle. \tag{9.59}$$

Next,

$$i \left\langle \frac{\partial B}{\partial t} \right\rangle = -\frac{3i}{2k_1} x^{-5/2} \dot{x} \langle p\phi + \phi p \rangle = -\frac{3i}{2} x^{-1} \dot{x} \frac{d\chi}{dt}. \tag{9.60}$$

Using these expressions we get

$$\frac{d^2\chi}{dt^2} + \frac{3}{2} \frac{\dot{x}}{x} \frac{d\chi}{dt} + \frac{\alpha}{x^3/2} \chi = \frac{2\omega}{k_1^2 x^3}. \tag{9.61}$$

We must now infer the behaviour of the quantum uncertainty χ in the limit of approach to the classically singular epoch using eqn (9.61). For this purpose, we first affect a change of variable from the time t to the scale factor Q; substituting $x = Q^2$, $\dot{x} = 2Q\dot{Q}$ in eqn (9.61), we obtain

$$\frac{d^2\chi}{dQ^2} \dot{Q}^2 + \frac{d\chi}{dQ} \left(\ddot{Q} + \frac{3\dot{Q}^2}{Q} \right) + \frac{\alpha\chi}{Q^3} = \frac{2\omega}{k_1^2 Q^6}. \tag{9.62}$$

The scale function Q here satisfies eqn (9.41) and, substituting for $\dot{Q}$ and $\ddot{Q}$ in the above, gives

$$\frac{d^2\chi}{dQ^2}(1-Q) + \frac{d\chi}{dQ} \frac{(5-6Q)}{2Q} + \frac{\chi}{Q^2} = \frac{2\omega}{\alpha k_1^2 Q^5}. \tag{9.63}$$

As yet, we have not made any assumption regarding the form of the wave function $\psi(\phi, t)$. To begin with, let us assume that ψ has a wave packet form peaked at $\phi = 0$ at the initial instance. Later we show that our results easily generalize to a completely general form of the wave function as well. Thus, if Δ is the quantum uncertainty in ϕ at the initial instance, we can write

$$\psi_i(\phi, t_i) = (2\pi\Delta_i)^{-1/4} \exp\left(\frac{-\phi^2}{4\Delta_i^2} \right). \tag{9.64}$$

In such a case, a simple calculation shows that $\chi\omega = 1/4$. Using this and a substitution $P = 1/Q$ in eqn (9.63) gives

$$\frac{\ddot{\chi}}{\chi}\left(1 - \frac{1}{P}\right) - \frac{1}{2P}\frac{\dot{\chi}}{\chi}\left(1 - \frac{2}{P}\right) + \frac{1}{P^2} = \frac{P}{2\alpha k_1^2 \chi^2}. \tag{9.65}$$

We are interested in the behaviour of χ in the limit of approach to the classical singularity, that is, in the limit as $Q \to 0$. In this limit, we can neglect the terms $1/P$, $1/P^2$ in eqn (9.65), because $P \to \infty$ in this situation. Then this equation reduces to

$$\frac{\chi\ddot{\chi}}{P} - \frac{1}{2}\frac{\chi\dot{\chi}}{P^2} = k, \qquad k > 0 \tag{9.66}$$

where $k = 1/2\alpha k_1^2$ and the dot denotes a differentiation with respect to P. Our problem now reduces to examining the behaviour of the non-linear differential equation (9.66). In fact it is possible to write down the solution in an implicit form which governs the evolution of χ. A new substitution $P = y^{2/3}$ reduces eqn (9.66) to the form

$$\chi\chi'' = A \tag{9.67}$$

where $A = \frac{4}{9}k > 0$ and the dash now denotes a derivative with respect to y. Integration of eqn (9.67) then gives

$$\frac{d\chi}{dy} = \pm(B\log\chi + C_1)^{1/2} = -\frac{2}{3}\frac{d\chi}{dQ}Q^{5/2} \tag{9.68}$$

with $B = 2A > 0$ and C_1 is a constant of integration such that $B\log\chi + C_1 > 0$.

At the initial epoch $Q = Q_0$, the collapse situation is highly classical and the quantum uncertainty χ around the classical state $\phi = 0$ must be small. However, as the collapse progresses and $Q \to 0$, we expect the quantum effects to grow since the action S will be of the order of the Planck constant h and in general χ increases. Thus χ is a decreasing function of Q and $d\chi/dQ$ should be negative. Then, writing $z = (B\log\chi + C_1)^{1/2}$ and integrating eqn (9.68) yields

$$Q^{-3/2} + C_2 = D\int e^{z^2/B}dz \tag{9.69}$$

where $D = 2e^{-C_1/B}/B > 0$. However, using a series expansion and integrating again, it is easy to see that

$$Q^{-3/2} + C_2 \leq \frac{2}{B}(B\log\chi + C_1)^{1/2}\chi. \tag{9.70}$$

Now in the limit of approach to the classical singularity $Q \to 0$, the left-hand side in (9.70) goes to infinity, which shows that χ cannot be finite in this limit and the quantum uncertainty must diverge as the collapse progresses.

It is not difficult now to generalize the above result to a situation when the wave function of the universe is no longer given in the wave packet form but is represented by a completely general wave function satisfying only conditions such as $\int \psi\bar{\psi} = 1$. In that case, we have $\chi\omega \geq 1/4$, and eqn (9.67) is modified to the inequality $\chi\chi'' \geq A$. Again following a similar procedure as above, we obtain

$$\frac{d\chi}{dQ} \geq \frac{-3}{2}(B\log\chi + C_1)^{1/2}Q^{-5/2} \tag{9.71}$$

and hence

$$Q^{-3/2} + C_2 \leq D\int e^{z^2/B}dz, \qquad D, B > 0. \tag{9.72}$$

It then follows that eqn (9.70) is again valid even when the quantum state of the universe is represented by a completely general wave function, and the uncertainty χ diverges in the limit of approach to the classical singular epoch.

Next, we examine the question of singularity within a fairly general cosmological scenario given by Belinskii, Lipschitz and Khalatnikov (1972) when quantum effects are important. The solution given by them contains eight arbitrary functions which describe a very general solution to Einstein's equations for a part of the space-time manifold. The solution here is given by

$$ds^2 = -dt^2 + (a^2 l_\alpha l_\beta + b^2 m_\alpha m_\beta + c^2 n_\alpha n_\beta)dx^\alpha dx^\beta, \tag{9.73}$$

where

$$a = t^{p_l}, b = t^{p_m}, c = t^{p_n}.$$

Here p_l, p_m, p_n and the vectors l, m, n are functions of space coordinates given by $\alpha, \beta = 1, 2, 3$ and satisfy

$$p_l + p_m + p_n = p_l^2 + p_m^2 + p_n^2 = 1. \tag{9.74}$$

Here, (9.74) implies that

$$\sqrt{-g} = tv \tag{9.75}$$

where the volume $v = l.(m \times n)$ is a function of space coordinates only.

We are basically interested in the behaviour in the vicinity of the classically singular epoch, where, as pointed out by Belinskii, Lipschitz and Khalatnikov, the contributions from the matter part L_m can be neglected. The approach to the singularity here consists of a succession of Kasner-type solutions where the space contracts along two directions and expands along the third one.

Thus, using $S_m = 0$ and eqn (9.75), the total action $S = S_m + S_H$ for the ensemble with conformal degree of freedom reduces to

$$S = -\frac{1}{16\pi G}\int_\nu 6\Omega_i\Omega^i\sqrt{-g}\,d^4x \tag{9.76}$$

where ν is the space-time region under consideration and $\Omega_i = \partial\Omega/\partial x^i$, the indices being raised and lowered by the classical metric g_{ij}. Then, integrating over the space coordinates in eqn (9.76), we obtain

$$S = \frac{3V}{8\pi G}\int_{t_1}^{t_2} t\dot{\phi}^2 dt, \tag{9.77}$$

where V is the 'volume' of the three-space under consideration.

This enables us now to write down the Hamiltonian for the system with the conformal degree of freedom. We assume that the state of the universe is represented by a wave function $\psi = \psi(\phi, t)$ and $\langle\phi\rangle = 0$, that is, the classical state $\phi = 0$ is always an average of the quantum ensemble.

We would like to examine the behaviour of the quantum uncertainty $\chi = \langle\phi^2\rangle$ in the limit of approach to the classically singular epoch.

From eqn (9.77), the Lagrangian is given by

$$L = (3V/8\pi G)\dot{\phi}^2 t, \tag{9.78}$$

and the Hamiltonian H as

$$H = \dot{\phi}p - L = p^2/2k_1 t, \qquad k_1 = 3V/4\pi G \tag{9.79}$$

where p is the associated conjugate momentum, $p = \partial L/\partial\dot{\phi}$. Choosing now $A = \phi^2$, the commutator relations as earlier and the evolution equation (9.44) give

$$d\chi/dt = (1/k_1 t)\langle p\phi + \phi p\rangle.$$

Our aim here is to generate a differential equation for the quantum uncertainty χ which indicates the possible departures from the classically singular state $\phi = 0$. For this we differentiate the above again, choosing

$$d\chi/dt = \langle B\rangle, \qquad B = (1/k_1 t)(p\phi + \phi p),$$

and using the commutators given by

$$[\phi, H] = ip/k_1 t, \qquad [p, H] = 0.$$

The evolution equation (9.44) then gives

$$\frac{d^2\chi}{dt^2} + \frac{1}{t}\frac{d\chi}{dt} = \frac{2\omega}{k_1^2 t^2} \tag{9.80}$$

where $\omega = \langle p^2 \rangle$ is the corresponding uncertainty in the momentum p.

Now, for a general wave function $\psi = \psi(\phi, t)$ for the universe, the quantum uncertainties χ and ω satisfy $\chi\omega \geq 1/4$. Treating the case $\chi\omega = \frac{1}{4}$ first, eqn (9.80) reduces to the following non-linear equation:

$$t^2\frac{d^2\chi}{dt^2} + t\frac{d\chi}{dt} = \frac{1}{2\chi k_1^2}.$$

We then use a new variable $P = 1/t$ to obtain

$$\chi\frac{d^2\chi}{dP^2} + \frac{\chi}{P}\frac{d\chi}{dP} - \frac{1}{2k_1^2 P^2} = 0 \tag{9.81}$$

In the limit of approach to the classical singularity, $t \to 0$ or $P \to \infty$, and the last term in eqn (9.81) can be neglected. It is then possible to solve the resulting equation when we note that χ must be small at the present epoch but must be larger in the vicinity of singularity where quantum effects are important and hence $d\chi/dt$ must be negative. We then obtain

$$\chi = -C\log t + C_1 \tag{9.82}$$

where C is a positive constant.

In the case of $\chi\omega > \frac{1}{4}$, in place of the above we have

$$t^2\chi'' + t\chi' > \frac{1}{2}\chi k_1^2,$$

where a prime denotes derivative with respect to t. In place of eqn (9.81) we then have

$$\chi\frac{d^2\chi}{dP^2} + \frac{\chi}{P}\frac{d\chi}{dP} > 0,$$

and a similar analysis leads to the inequality

$$\chi > -C\log t + C_1.$$

The above results show that even when the wave function of the universe has a completely general form, the quantum uncertainty χ must diverge logarithmically in the limit of approach to the classically singular epoch. Now, χ indicates the spread of the wave function around the classical state

$\phi = 0$. Hence, our results point out that within the quantum framework considered here, the singularity is no longer an inevitability and that non-classical, non-singular states occur with finite probability.

The quantum principles used here are again of general nature, namely, that the quantum state of the universe is represented by a general wave function $\psi(\phi, t)$ and that the time evolution of operators is given by eqn (9.44). Even though we have written down the Hamiltonian for the system with only the conformal degree of freedom, the divergence of the spread of the wave function illustrates that the classical singularity is no longer inevitable when quantum effects are taken into account. Thus, it can be stated with some degree of confidence that quantum cosmology offers a hope of alleviating the classical singularity problem.

To discuss a general space-time now, the metric in the vicinity of a spacelike hypersurface, for any stably causal or globally hyperbolic space-time, can be written as

$$ds^2 = -dt^2 + g_{\alpha\beta} dx^\alpha dx^\beta, \tag{9.83}$$

where $\alpha, \beta = 1, 2, 3$ denote the space coordinates.

We can foliate the space-time by means of a sequence of spacelike hypersurfaces Σ_t and without loss of generality we can take the singularity to be at $t = 0$. We can also consider the case when the space-time contains singularities arising out of local collapse, in addition to global cosmological singularities, and our considerations can be generalized to include this situation as well. Then the action is written as

$$S = \frac{1}{16\pi} \int_\nu [(1 + \phi^2) h_1(t) - \dot{\phi}^2 h_2(t)] dt, \tag{9.84}$$

where the dot denotes differentiation with respect to time and we have introduced the notation

$$h_1(t) = \int_{\nu \cap \Sigma_t} \bar{R} \sqrt{-\bar{g}}\, d^3x,$$

$$h_2(t) = 6 \int_{\nu \cap \Sigma_t} \sqrt{-\bar{g}}\, d^3x.$$

To begin with, we discuss the class of spatially homogeneous models and take $\phi = \phi(t)$ only to be consistent with the symmetry of the classical problem. Here Σ_t is a member of the sequence of the Cauchy surfaces or partial Cauchy surfaces approaching the singularity. The functions $h_1(t)$ and $h_2(t)$ represent spatial integration carried out on a finite region of space. We will need to discuss the behaviour of these functions later. Then, from eqn (9.84), the Lagrangian for the ensemble with the conformal degree of freedom can be written as

$$L = \frac{1}{16\pi}[(1+\phi^2)h_1(t) - \dot{\phi}^2 h_2(t)]. \tag{9.85}$$

The associated conjugate momentum is given by

$$p = \frac{\partial L}{\partial \dot{\phi}} = -\frac{1}{8\pi}\dot{\phi}h_2(t). \tag{9.86}$$

We can now write down the Hamiltonian for the system:

$$H = \dot{\phi}p - L = -\frac{4\pi p^2}{h_2(t)} - \frac{1}{16\pi}(1+\phi^2)h_1(t). \tag{9.87}$$

In the classical scenario the state $\phi = 0$ occurs with certainty, which corresponds to the space-time manifold with a curvature singularity. Quantum mechanically, however, there will be a dispersion around the value $\phi = 0$ and states with a non-zero value of ϕ can occur. The dispersion from the value $\phi = 0$ basically denotes the quantum effects in our considerations. We define the dispersions in the variables ϕ and p as earlier

$$\chi = \langle\phi^2\rangle - \langle\phi\rangle^2, \quad \omega = \langle p^2\rangle - \langle p\rangle^2.$$

In particular, we would like to derive the behaviour of χ in the limit of approach to the classical singularity and then interpret the same.

Let now $\psi(\phi, t)$ be a general wave function representing the state of the system. As earlier we assume that the classical state $\phi = 0$ remains, at all times the average for the quantum description; that is, at any time t,

$$\langle\phi_t\rangle = \int_{-\infty}^{\infty} \bar{\psi}(\phi_t)\phi_t\psi(\phi_t)d\phi_t = 0. \tag{9.88}$$

Then we have

$$0 = \frac{d}{dt}\langle\phi\rangle = i\int \bar{\psi}[H\phi - \phi H]\psi dt,$$

which gives, when we use the operator representation $p = (1/i)(\partial/\partial\phi)$,

$$0 = -\frac{8\pi\langle p\rangle}{h_2(t)}. \tag{9.89}$$

Thus, $\langle p\rangle = 0$ at all times and we have

$$\chi = \langle\phi^2\rangle, \quad \omega = \langle p^2\rangle.$$

We note that the classical metric (9.83), obtained as a solution to the Einstein equations, represents the singular situation corresponding to $\phi = 0$. The other states in the quantum ensemble need not satisfy Einstein equations and could well be non-singular. Defining an operator by $A = \phi^2$, and $\chi = \langle A \rangle$ the quantum effects χ can be evolved in time using the evolution equation (9.44). Successive applications of (9.44) yield a non-linear differential equation for χ which can be analysed to deduce its behaviour. (This will be an inequality for the case of a general wave function satisfying $\chi\omega \geq 1/4$.) Using the Hamiltonian H, the commutators $[\phi, H]$ and $[p, H]$ are given as

$$[\phi, H] = i\frac{\partial H}{\partial p} = -i\frac{8\pi p}{h_2(t)},$$

$$[p, H] = -i\frac{\partial H}{\partial \phi} = i\frac{\phi h_1(t)}{8\pi}.$$

Then

$$[\phi^2, H] = -i\frac{8\pi}{h_2(t)}(p\phi + \phi p),$$

and we get

$$\frac{d\chi}{dt} = -\frac{8\pi}{h_2(t)}\langle p\phi + \phi p \rangle.$$

Defining a new operator B by $\langle B \rangle = d\chi/dt$ we again use eqn (9.44). We have

$$[B, H] = -\frac{8\pi}{h_2(t)}[p\phi + \phi p, H],$$

and again using the commutator relations above this can be evaluated to be

$$[B, H] = -\frac{2ih_1(t)}{h_2(t)}\phi^2 + \frac{128\pi^2 i}{h_2^2(t)}p^2. \tag{9.90}$$

Next, we get

$$\frac{\partial B}{\partial t} = -\frac{\dot{h}_2(t)}{h_2(t)}\frac{d\chi}{dt}.$$

Using the above relations in the evolution equation gives the required differential equation for χ :

$$\frac{d^2\chi}{dt^2} + \frac{\dot{h}_2}{h_2}\frac{d\chi}{dt} + \frac{2h_1}{h_2}\chi = \frac{128\pi^2\omega}{h_2^2}. \tag{9.91}$$

Again, the quantum state of the universe is represented by a general wave function $\psi = \psi(\phi, t)$ and χ and ω satisfy $\chi\omega \geq 1/4$. When ψ has the wave packet form, the equality holds in the above. First we assume equality which is easily generalized as we shall point out later. It is convenient to write $t = 1/r$ to get

$$
\begin{aligned}
h_1(t) &= h_1(1/r) = g_1(r), \\
h_2(t) &= h_2(1/r) = g_2(r), \\
\dot{h}_2(t) &= g_2'(r)\frac{dr}{dt} = -r^2 g_2'(r),
\end{aligned}
$$

where a dash denotes differentiation with respect to r. With this substitution, eqn (9.91) is written as

$$
g_2^2\frac{\chi''}{\chi} + g_2^2\left(\frac{2}{r} + \frac{g_2'}{g_2}\right)\frac{\chi'}{\chi} + \frac{2g_1g_2}{r^4} = \frac{32\pi^2}{\chi^2 r^4}. \tag{9.92}
$$

It is now necessary to consider the nature of the functions $h_1(t)$ and $h_2(t)$ as defined above. For the Friedmann–Robertson–Walker case considered earlier, they have the form,

$$
\begin{aligned}
h_1(t) &= 3V\ddot{x}x^{1/2} = 6VQ(\dot{Q}^2 + Q\ddot{Q}), \\
h_2(t) &= 6VQ^3.
\end{aligned}
$$

For $k = 0$ universe, $Q \sim t^{2/3}$ and we get $h_1(t)$ to be a constant and $h_2(t) \sim t^2$. Thus, the product $h_1(t)h_2(t)$ vanishes in the limit of approach to the singularity. A similar behaviour will be seen for the case of a collapsing dust ball to be discussed later. In fact, the function $h_2(t)$ is a special case of the volume element $V(t)$ defined in Section 5.4 along a timelike geodesic. This is because the three-form $\mu(t)$ can be regarded as the metric determinant when defined using a geodesically comoving coordinate system. Thus, $h_2(t) \to 0$ in the limit of approach to the singularity at $t = 0$. Again, we expect $h_1(t)$ to remain at least finite in this limit. Hence, we assume that the term $2g_1g_2/r^4$ vanishes in the limit as $t \to 0$ and we ignore the same. The resulting equation from (9.92) can be simplified considerably if we define a new variable τ by

$$
\tau = -\int \frac{dr}{r^2 g_2(r)}, \tag{9.93}
$$

and this non-linear differential equation reduces to the form

$$
\chi\frac{d^2\chi}{d\tau^2} = A, \qquad A > 0. \tag{9.94}
$$

Here, $A = 32\pi^2$. Equation (9.94) again is a non-linear equation; however, it is possible to analyse the same using a method similar to that used earlier. We can integrate to obtain

$$h_2(t)\frac{d\chi}{dt} = \pm(2A\log\chi + C_1)^{1/2}. \tag{9.95}$$

At the initial epoch $t = t_1$ the situation is highly classical and the quantum uncertainty χ around the classical state must be very small. However, as $t \to 0$, the action S is of the order $\hbar$ and in general χ increases. Thus, χ is a decreasing function of t and we choose the negative sign in eqn (9.95). Writing $z = (2A\log\chi + C_1)^{1/2}$ and using a series expansion for e^{z^2} one can again integrate the above to obtain finally

$$\left|\int \frac{dt}{h_2(t)}\right| \leq (A\log\chi + C_1)^{1/2}\chi. \tag{9.96}$$

As pointed out above, clearly $h_2(t) \to 0$ in the limit of approach to the classically singular epoch. The integral on the left will diverge if $h_2(t)$ goes to zero at least as far as t. For the case of homogeneous and isotropic cosmologies, $h_2(t) \sim t^2$ as seen above. Further, for all known important examples of exact solutions such as the collapsing dust ball, Bianchi models, or the general cosmological scenario of Belinskii *et al.* this condition is satisfied. Finally, we considered in Section 5.4 a general class of strong curvature singularities where $F(t) \sim 1/t^2$ and it was shown there that $V(t)$, which is proportional to $h_2(t)$, goes as $t^{3/2}$ near the singularity. Thus, we conclude that for a very wide class of physically reasonable strong curvature singularities, the integral on the left of eqn (9.96) must diverge. Hence, the quantum uncertainty χ must diverge in the limit of approach to the classical singularity.

This result shows that the quantum effects, which were negligible initially, grow and become very important in the vicinity of space-time singularity. Finally, we note that it is not difficult to generalize to the situation when ψ has a general form and $\chi\omega \geq 1/4$. Then, in place of eqn (9.94), we get

$$\chi d^2\chi/d\tau^2 \geq A.$$

Then a similar procedure can again be carried out to arrive at eqn (9.96).

It is useful to indicate the relationship of the above result with the case of homogeneous dust collapse discussed above. In this case, the functions $h_1(t)$ and $h_2(t)$ are then worked out to be

$$h_1(t) = 8\pi\rho_0 V,$$

$$h_2(t) = 6Q^3V.$$

In fact $Q = \frac{3}{2}\alpha t^{2/3}$ here and we again get $h_2(t) \sim t^2$ in the limit as $t \to 0$. Substituting this in eqn (8.46) and writing $x = Q^2$ we get

$$\frac{d^2\chi}{dt^2} + \frac{3}{2}\frac{\dot{x}}{x}\frac{d\chi}{dt} + \frac{\alpha}{x^{3/2}}\chi = \frac{2\omega}{k_1^2 x^3}, \tag{9.97}$$

where $k_1 = -3V/4\pi$. Thus we see that the case of homogeneous collapse is contained as a special case in the present results.

Throughout we have presented the calculations with the assumption $\phi = \phi(t)$. However, it is not difficult to generalize the conclusions to the situation when the fluctuation ϕ is both a function of space and time coordinates and we present below an outline of the argument. So far the expectation values $\bar{\chi} = \langle\bar{\phi}^2\rangle$ are evaluated over all possible configurations (constants) and given by

$$\langle\bar{\phi}^2\rangle = \int_{-\infty}^{\infty} \psi\bar{\phi}^2\bar{\psi}d\bar{\phi},$$

where the values $\bar{\phi}$ belong to the real line. Suppose now, for example, that the state of the system is described by a discrete set of variables $\phi^{(1)}, \phi^{(2)}, \ldots, \phi^{(n)}$ rather than a single variable $\bar{\phi}$. The average value of an observable F at a time t is then given by

$$\langle F\rangle = \int \cdots \int \psi(\phi^{(1)}, \ldots, \phi^{(n)})F\bar{\psi}(\phi^{(1)}, \ldots, \phi^{(n)})d\phi^{(1)}\ldots d\phi^{(n)}. \tag{9.98}$$

Now, for a transition to the situation $\phi = \phi(x^i)$, a continuum limit is to be taken in the above, as is normally done in evaluating path integrals (Feynman and Hibbs, 1965), and we have

$$\langle F\rangle = \int \psi(x^i)F\bar{\psi}(x^i)\mathcal{D}\phi(x^i). \tag{9.99}$$

Here $\mathcal{D}$ denotes a measure on the space of all functions ϕ on a given hypersurface Σ_t. We can think of $\phi(x^i)$ as a 'field' with infinitely many degrees of freedom. The infinitely many numbers $\phi(x^i)$, obtained by varying x^i, characterize the state of the system.

We have earlier evaluated the action S and defined the functions $h_1(t)$, $h_2(t)$, and so on, by integrating over a finite range of variables in space. This was done at a constant time t, integrating over a compact region ν in the spacelike hypersurface Σ. A limit was taken then, as the sequence of hypersurfaces approached the singular epoch. Define now

$$\bar{\phi} = \min_{\Sigma_t \cap \nu}\phi(t, x^\alpha), \quad \alpha = 1, 2, 3.$$

Evaluating such minimum over all possible functions at a given time t, we have an entire range of constants given by $-1 < \phi < \infty$. Clearly, we always have,

$$\bar{\phi} \leq \phi(x^i).$$

When the state is described by a finite set of variables $\phi^{(1)}, \phi^{(2)}, \ldots, \phi^{(n)}$ then it is clear from eqn (9.98) that whenever $F \leq G$ for any two functions F and G, then $\langle F \rangle \leq \langle G \rangle$. In the continuum limit given by eqn (9.99), there is generally an ambiguity concerning the question of measure. However, in such a case also, we would expect that for any reasonable definition of the measure $\mathcal{D}$ we must have $\langle F \rangle \leq \langle G \rangle$ whenever $F \leq G$ pointwise. Then because $\bar{\phi}^2 \leq \phi^2$, we obtain $\bar{\chi} \leq \chi$ where $\bar{\chi}$ and χ are the expectation values of $\bar{\phi}^2$ and ϕ^2 respectively. Since we have shown earlier that $\bar{\chi}$ diverges in the limit of approach to the classical singularity, it follows that χ also must diverge in the same limit. This recovers the earlier result on quantum effects for the case when $\phi = \phi(x^i)$.

9.5 Probability measure for singular geometries

Using the results developed so far, we have been able to draw the conclusion that the quantum conformal fluctuations diverge near the epoch where the classical geometry admits a space-time singularity. This means that the classical picture can no longer be taken as a representative near this epoch and conclusions concerning the 'inevitable' occurrence of singularities have to be revised. The quantum conformal framework may admit geometries which are no longer infected with the singularity problem. However, this raises the important question mentioned earlier: given the full range of possible geometries within the quantum conformal framework, what is the probability measure of the class of geometries which are singular at the classically singular epoch? If it turns out that after quantizing the conformal fluctuations the set of singular geometries has a non-zero measure of probability, then it means that the singularity problem has not been avoided in the new framework. On the other hand, if the probability measure for the occurrence of singular solutions were to have a zero measure, we would have effectively eliminated the singularity problem by quantizing the conformal fluctuations.

Again we begin by analysing the simple situation of homogeneous and isotropic space-times where ϕ depends on t and the quantum state of the universe is described by a completely general wave function:

$$\Psi = \Psi(\phi, t, \triangle). \tag{9.100}$$

We have singled out $\triangle$ above to indicate the dispersion of Ψ. Let the singularity be denoted by S, which is characterized by the TIP $I^-(S)$ and generated by some future endless non-spacelike curve γ of finite length. Recalling the results from Section 9.2, we see that as $X \to S$ along γ, if ϕ behaves as $\phi = K\phi_c$ with $\phi_c = \Omega_c - 1$, and $K > 0$ is a constant, then ϕ blows up sufficiently fast along γ and the lengths of all non-spacelike curves in $I^-(S)$ can no longer be finite, which is a contradiction. Thus, for the singularity S to be admitted in M, we must have as $X \to S$,

$$\phi < K\phi_c, \qquad \text{for all} \quad K > 0, \tag{9.101}$$

that is, $\phi/\phi_c \equiv \alpha$ must tend to zero.

Defining a new variable $u = \phi/\Delta$ if we now rescale Ψ as

$$\psi(u,t) = \sqrt{\Delta}\Psi(u\Delta, t, \Delta), \tag{9.102}$$

then the new wave function $\psi(u,t)$ is seen to have a unit dispersion at all times and the condition for singularity becomes

$$\frac{u\Delta}{\phi_c} \to 0. \tag{9.103}$$

Thus the singular solutions are defined by those values of u such that

$$u = \alpha\frac{\phi_c}{\Delta}, \quad \alpha \to 0. \tag{9.104}$$

Now, as shown in Section 9.2, $\Delta \sim A_{IF}^{-1}$. Further, eqns (9.11) and (9.37) show that A_{IF}, which is the inverse of the Greens function G goes as ϕ_c^{-1}. Thus we have $\Delta \sim \phi_c$ in this case and eqn (9.104) becomes

$$u < \alpha\beta, \tag{9.105}$$

where $\beta =$ const. and $\alpha \to 0$. Since the probability density is given by

$$\mid \psi(u,t) \mid^2 = \bar{\psi}(u,t)\psi(u,t),$$

the probability measure of the set of geometries which will be singular, is given by

$$P_s = \int_{|u|<\alpha\beta} \mid \psi(u,t) \mid^2 du, \qquad \alpha \to 0. \tag{9.106}$$

The only circumstance under which P_s would remain non-zero in this limit is if $\mid \psi \mid^2$ approaches a multiple of $\delta(u)$. However, this is impossible since the dispersion of ψ is unity—unless the limiting form of ψ is such as to give for $\mid \psi \mid^2$ a delta function superposed on a function of support extending beyond $| u |= 1$. Excluding such pathological cases, we therefore conclude that among all possible geometries, the singular geometries will occur only with a vanishing probability. Note also that the geometries in the non-singular class defined by $| u |> \alpha\beta$ contribute predominantly to the probability measure as $X \to S$. We have thus shown that within the quantum conformal framework the measure of the set of those geometries which are likely to occur with singularities is zero.

Next, let us consider the classical situation of a general globally hyperbolic space-time with arbitrary distribution of material particles and many singularities spread over at various epochs. One could also have here all embracing cosmological singularities in the future or past in the case of 'closed' universes, that is, those which admit a compact Cauchy surface. A globally hyperbolic M admits what is called a 'cosmic time function' which is a real-valued function f from $M \to \mathbb{R}$ such that f always increases along any future directed non-spacelike curve and ∇f is always future pointing and timelike. We can foliate M using a Cauchy cosmic time function which has the property that for each $c \in \mathbb{R}, f^{-1}(c)$ is a Cauchy surface in M. Thus, the cosmic time function provides a choice of foliation which we shall use here.

Let $S_1, S_2, \ldots, S_n \ldots$ be the space-time singularities in M which are characterized by the TIP's $I^-(S_1), I^-(S_2), \ldots, I^-(S_n) \ldots$, and so on, generated by future directed non-space-like curves γ_1, γ_2, and so on, that have finite lengths in future but no future end points. Now, consider a singularity S_i lying on a spacelike hypersurface Σ_{t_i}. Let the TIP $I^-(S_i)$ be generated by γ_i. As shown above, as $X \to S_i$ along γ_i, we must have $\phi/\phi_c = \alpha \to 0$, and so the condition for the geometry to be singular at an epoch $t = t_i$ is given as

$$u < \alpha \frac{\phi_c}{\Delta} \quad \text{along} \quad \gamma_i. \tag{9.107}$$

Now, in the above general situation under consideration, the state of the universe at any epoch t is no longer described by a single variable $\phi = \phi \in \mathbb{R}$ as in the homogeneous, isotropic case, but is described by an infinite set of continuum variables $\phi = \phi(x^\mu, t)$. In this situation, the evolution of the wave function is given by eqn (9.13) where the propagator in this general case is given by eqn (9.9). We adopt here the following simplifying notation used in Narlikar (1984). A repeated continuous variable J in any expression would mean an integration over the entire range. Thus, the propagation equation for a wave function can be expressed as

$$\Psi_F(\phi_F) = K[\phi_F, t_F; \phi_I, t_I]\Psi_I(\phi_I).$$

With this notation, the general propagator (9.9) can be written as

$$K[\phi_F, t_F; \phi_I, t_I] = F(t_F, t_I) \exp i[A_{FF}(X_F, X_F')\phi_F(X_F)\phi_F(X_F') + A_I(X_I)\phi_I(X_I)^2 + 2A_{IF}(X_I, X_F)\phi_I(X_I)\phi_F(X_F)] \tag{9.108}$$

where, in the second term, we have diagonalized the quadratic form

$$A_{II}\phi_I(X_I)\phi_I(X_F').$$

Now, suppose the wave functional describing the state of the universe is given by

$$\Psi_I[\phi_I(X)] = \frac{1}{[2\pi\Delta_I(X)]^{1/4}} \exp\left\{-\frac{[\phi_I(X)]^2}{4[\Delta_I(X)]^2}\right\}.$$

The approximation that the present state of the universe is almost classical means that $|\Delta_I| \ll 1$. Using eqn (9.13) and the above we can evolve Ψ_I to the wave functional Ψ_F at any other later epoch $t = t_F$ to get that $\Psi_F[\phi_F(X)]$ is proportional to

$$\exp\left(\frac{A_{IF}^2\phi_I(X)^2}{(iA_I - 1/4\Delta_I^2)} + iA_{FF}(X, X')\phi_F(X)\phi_F(X')\right).$$

Here, the normalizing constant of proportionality does not depend on ϕ_I. Now, the above expression can be used to obtain the probability density

$$|\Psi_F|^2 = f \exp\left(-\frac{{\phi_F}^2}{2\Delta_F^2}\right), \quad f = \text{const.},$$

where, in analogy to eqn (9.17),

$$\Delta_F^2 = \Delta_I^2\left(A_I^2 + \frac{1}{16{\Delta_I}^4}\right)A_{IF}^{-2}.$$

Thus, we see that as $X \to S_i$, $\Delta \sim A_{IF}^{-1}$. However, as seen earlier,

$$A_{IF} \sim \bar{G}(X_2, X_1)^{-1} \sim \Omega_c^{-1},$$

and hence we have $\Delta \sim \phi_c$. As a result, we can write

$$\Delta = \sigma(X^\mu, t)\phi_c, \tag{9.109}$$

where σ is a well-behaved function along γ_i. Thus, the condition (9.107) for the singularity now becomes

$$u < \alpha\beta(X^\mu, t). \tag{9.110}$$

Now we can write the probability measure for the class of geometries in the set of all possible geometries which will be singular at $t = t_i$, as

$$P_S = \int_{|u|<\alpha\beta} |\ \Psi(u,t)\ |^2\ Du = \int_{|u|<\alpha\beta} f \exp(-u^2/2) Du.$$

Clearly, the above integral becomes arbitrarily small in the limit as $\alpha \to 0$ and hence we deduce that in the limit of approach to the classically singular epoch $t = t_i$ it is vanishingly small.

Finally, it is not difficult to see that the analysis such as above can be extended to wave functionals and analogous conclusions drawn for Ψ_I not necessarily described as wave packets. It is thus seen that within the set of all geometries conformal to a given singular classical geometry, the subset of singular geometries occurs with zero probability. The probability in this work is defined in terms of conformal quantization of Einstein's equations.

This work, of course, does not deal with non-conformal degrees of freedom, and to this extent the overall question of how probable singularities are in general quantum cosmology remains open. Nevertheless, the direct connection between the conformal transformations and singularities discussed here allows for the possibility that singular geometries may turn out to be exceptions rather than the rule in a full theory of quantum gravity.

9.6 Quantum effects near a black hole singularity

A well-known example of quantum effects in a black hole geometry is the result that the black hole must radiate thermally (Hawking, 1975). The hope is that such results may form part of the full theory and may eventually even lead us to the full theory.

We would like to examine here the evolution of quantum effects within a black hole geometry using the approach being used in this chapter which is somewhat similar to the minisuperspace approach where one quantizes a limited degree of freedom for the metric tensor. We quantize the conformal freedom Ω for the classical black hole metric g_{ij}, where the conformal fluctuation is given as $g_{ij} \to \Omega^2 g_{ij}$. Usually the quantum effects can distort the underlying space-time geometry and results may become ambiguous. However, in the present approach the causal structure of the space-time is left invariant by the quantum fluctuations and the results become much more transparent. We shall, in fact, show here that the dispersion $\chi = \langle \phi^2 \rangle$ (where $\Omega = 1 + \phi$) diverges near the singularity and hence quantum effects dominate. Possible interpretations and implications of the above result are briefly indicated.

The geometry of a spherically symmetric static black hole, which would result for example from a homogeneous gravitational collapse of dust to the centre $r = 0$, is given in the coordinates (t, r, θ, ϕ) by the usual Schwarzschild metric,

$$ds^2 = -\left(1 - \frac{2m}{r}\right) dt^2 + \left(1 - \frac{2m}{r}\right)^{-1} dr^2 + r^2(d\theta^2 + \sin^2\theta d\phi^2). \quad (9.111)$$

The form (9.111) has, however, a coordinate singularity at $r = 2m$ and it covers the manifold only in the range $2m < r < \infty$. As we want to examine the effects in the vicinity of the singularity $r = 0$, it is necessary to work in the Kruskal space-time (see Fig. 13) which covers the full region $r > 0$ and is given by

$$ds^2 = \frac{32m^3}{r} \exp\left(\frac{-r}{2m}\right)(-dT^2 + dX^2) + r^2(d\theta^2 + \sin^2\theta\, d\phi^2). \quad (9.112)$$

Here the coordinate r is defined implicitly by the relation

$$X^2 - T^2 = \left(\frac{r}{2m} - 1\right) \exp\left(\frac{r}{2m}\right) \quad (9.113)$$

as explained in Section 3.2. We would like to examine the ensemble $\Omega^2 ds^2$ for all possible Ω with $0 < \Omega < \infty$. The classical action S for the gravitational field then transforms to a new value as given earlier. For simplicity of presentation, we will choose here $\phi = \phi(T)$ only; the general case $\phi = \phi(X, T)$ can be dealt with without much further complication in a similar way. Then, using $g^{00} = -(r/32m^3)\exp(r/2m)$ and $R = 0$, since the exterior space-time is empty, we can write for the action

$$S = \int_{\mathcal{V}} 6\dot{\phi}^2 r^2(X, T) \sin\theta\, d^4x. \quad (9.114)$$

Here the dot denotes a differentiation with respect to T and $\phi = \Omega - 1$. Thus the Lagrangian for the system with conformal degree of freedom can be written as

$$L = k\dot{\phi}^2 S(T) \quad (9.115)$$

where $k = 24\pi > 0$ is the result of integration over the angular coordinates and $S(T) = \int r^2(X, T) dX$. Defining the conjugate momentum as $p = \partial L / \partial \dot{\phi}$, the Hamiltonian turns out to be

$$H = \dot{\phi} p - L = p^2 / 4kS(T) \quad (9.116)$$

In order to quantize, we treat now ϕ and p as operators. We note that $\phi = 0$ gives us back the classical black hole space-time. We shall assume that, even when quantum fluctuations take place, the classical state still remains an average of the quantum ensemble, that is, $\langle\phi\rangle = 0$ at all times. Then, using

$$0 = \frac{d}{dt}\langle\phi\rangle = i\int \bar{\Psi}[H\phi - \phi H]\Psi d\tau \tag{9.117}$$

and the operator representation $p = (1/i)(\partial/\partial\phi)$, we get $\langle p\rangle = 0$. In the above, $\Psi = \Psi(\phi, T)$ is the wave function specifying the state of the system. We shall not need here to make any specific assumption regarding the form of Ψ, such as a Gaussian wave packet form, which we will take to be any general function satisfying conditions such as $\int \Psi\bar{\Psi} = 1$, and so on. Then, defining $\chi = \langle\phi^2\rangle - \langle\phi\rangle^2$ and $\omega = \langle p^2\rangle - \langle p\rangle^2$, we have a general inequality such as $\chi\omega \geq 1/4$. The time evolution of the operator $A = \phi^2$ is governed by eqn (9.44). Now working out the commutators such as $[\phi, H], [p, H], [\phi p, H]$, and applying eqn (9.44) twice, yields the following non-linear differential equation for χ:

$$\chi\chi'' + \left(\frac{2}{q} + \frac{h_1'(q)}{h_1(q)}\right)\chi\chi' = \frac{1}{8q^4 h_1^2(q)k^2} \tag{9.118}$$

In the above, we have used $T = 1/q$ and $S(T) = S(1/q) = h_1(q)$, and a prime denotes a differentiation with respect to q. Also, we have used $\chi\omega = 1/4$ to eliminate ω. Again, without much difficulty this situation can be generalized to $\chi\omega \geq 1/4$ for the case of a general wave function. Whereas for the first application of eqn (9.44), $\partial A/\partial T = 0$, for the second application, the operator $B = (p\phi + \phi p)/2kS(T)$ turns out to have explicit time dependence, which is to be taken into account while doing the above calculation. A substitution given by

$$\tau = -\int \frac{dq}{q^2 h_1(q)}, \tag{9.119}$$

reduces eqn (9.118) to a remarkably simple form given by

$$\chi\frac{d^2\chi}{d\tau^2} = k_1^2, \qquad k_1 = \frac{1}{8k^2}. \tag{9.120}$$

It is now possible to integrate the above equation by means of usual techniques and one obtains

$$\frac{d\chi}{dT} = \pm\frac{(C_1 \log\chi + C_2)^{1/2}}{S(T)}, \qquad C_1 > 0 \tag{9.121}$$

We shall choose here the positive sign for $d\chi/dT$ because one expects the quantum effects to grow generally in the neighbourhood of a singularity at $r = 0$, whereas these must be negligible far away from the singularity. However, even if $d\chi/d\tau < 0$ initially, say at $\tau = \tau_0$, this derivative would still become positive at a later epoch provided χ is bounded below by a positive quantity at this initial epoch. To see this, write the integral of eqn (9.120) as $(d\chi/d\tau)^2 = A\log\chi + C$. Then, if $d\chi/d\tau < 0$ and $\chi > e^{-C/A}$ initially, differentiating the above first integral gives

$$\frac{d^2\chi}{d\tau^2} = \frac{A}{\chi} > \frac{A}{\chi(\tau_0)},$$

for any $\tau > \tau_0$ with $d\chi/d\tau < 0$ between τ_0 and τ. Then integrating the above equation in this region,

$$\frac{d\chi}{d\tau} > \left(\frac{d\chi}{d\tau}\right)_{|\tau=\tau_0} + \frac{A(\tau - \tau_0)}{\chi(\tau_0)}.$$

This implies that $d\chi/d\tau$ must vanish at some value of τ less than

$$\tau_0 - \left(\frac{d\chi}{d\tau}\right)_{|\tau=\tau_0} \frac{\chi(\tau_0)}{A}.$$

After this value the derivative again becomes positive. In such a case, the limit of approach to the singularity is described by the Kruskal picture given in Chapter 3 by increasing values of the coordinate T in the black hole region II and hence χ must increase with increasing values of T.

It is now possible to integrate eqn (9.121) again by using a new substitution $z = (C_1 \log\chi + C_2)^{1/2}$ and using the series expansion for $\exp(z^2/C_1)$. The result is the following inequality relating $S(T)$ and χ:

$$\int \frac{dT}{S(T)} \leq \frac{2}{C_1}(C_1 \log\chi + C_2)^{1/2}\chi. \tag{9.122}$$

Our task now is to analyse this inequality in order to deduce the behaviour of χ in the vicinity of the black hole singularity. Using eqn (9.113) and a series expansion for $\exp(r/2m)$ it is not difficult to see that, when $r \geq 2m$, the following inequality holds:

$$r^2(X,T) \leq 4m^2(3X^2 - T^2 + 1). \tag{9.123}$$

For $0 < r < 2m$, we are in the region II or the black hole region of the Kruskal picture (see Fig. 60) and $-1 < X^2 - T^2 < 0$. The event horizon at

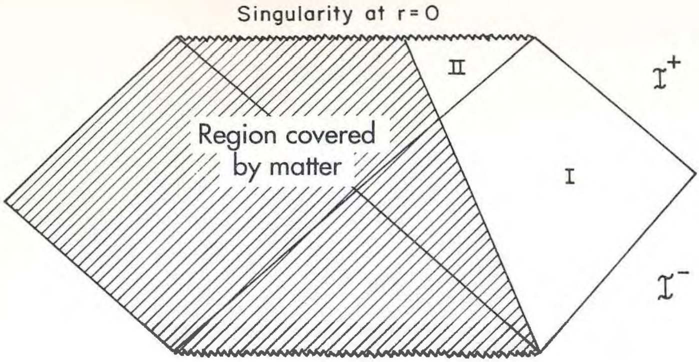

Fig. 60 The shaded part of the Kruskal picture is the region covered by the collapsing matter.

$r = 2m$ is then given by $X = T$ and the singularity at $r = 0$ is represented by the curve $X^2 - T^2 = -1$. It is again seen by means of a simple calculation that the condition $X > \sqrt{1/2}$ implies that eqn (9.123) holds in this situation. This condition on the X coordinate means that some part of the singularity curve is covered by the collapsing matter, so the X coordinate should have a lower bound. This will be determined purely by the geodesic structure of the trajectories of the collapsing material particles.

Any point on the singularity curve is given by the two coordinates (X_s, T_s) governed by the condition $X_s^2 - T_s^2 = -1$. In order to examine the behaviour of χ near singularity, we have to study eqn (9.122) as this limit is approached. For this, first $S(T)$ is to be evaluated which is then integrated from some initial epoch T_0 to the singularity T_s. Then, using eqn (9.123),

$$\begin{aligned} S(T) &= \int_{X_a(T)}^{X_s} r^2(X,T)dX \leq 4m^2 \int_{X_a(T)}^{X_s} (3X^2 - T^2 + 1)dX \\ &= 4m^2[X^3 - T^2X + X]_{X_a(T)}^{X_s}. \end{aligned} \tag{9.124}$$

However, for $r > 0$, we have $X^2 - T^2 > -1$ and this gives

$$4m^2 X_a[X_a^2(T) - T^2 + 1] > 0,$$

because $X_a > 0$. Hence we get

$$S(T) < 4m^2 X_s(X_s^2 - T^2 + 1).$$

Then, going back to eqn (9.122) gives

$$\frac{1}{4m^2X_s}\int_{T=T_0}^{T_1}\frac{dT}{X_s^2+1-T^2}=\frac{1}{8m^2X_s(X_s^2+1)^{1/2}}\log\left[\frac{(X_s^2+1)^{1/2}+T}{(X_s^2+1)^{1/2}-T}\right]_{T_0}^{T_1}$$
$$\leq \chi(C_1\log\chi+C_2)^{1/2}. \tag{9.125}$$

Now, in the limit of approach to the singularity,

$$T_1 \to T_s = (X_s^2+1)^{1/2},$$

and hence the left-hand side of eqn (9.125) diverges. Thus, in the limit of approach to the classical singularity we obtain the result that $\chi \to \infty$, that is, the quantum uncertainty diverges.

Even though classically the singularity is the only possible final state for the black hole space-time, we see that quantum mechanically states with arbitrary large ϕ can occur. As pointed out in earlier sections, such space-times need not necessarily be singular. We might thus have black holes without singularity in the following sense: the gravitational collapse may proceed classically to form the event horizon and hence the black hole. However, quantum effects take over then and the singularity is no longer a unique outcome; in fact a non-singular geometry with large ϕ can occur with a finite probability. Such ideas of having a black hole without a singularity have been explored in the classical context as well (Newman, 1989).

REFERENCES

Abbott L. F. and Sikivie P. (1983). 'A Cosmological Bound on the Invisible Axion', Phys. Lett. 120B, p. 133.

Akolia G. M. , Joshi P. S. and Vyas U. D. (1981). 'On Almost Causality', Jour. Math. Phys. 22, p. 1243.

Alexandrov A. D. (1950). 'On Lorentz Transformation', Usp. Mat. Nauk. 5, p. 187.

Alexandrov A. D. (1967). 'A Contribution to Chronogeometry', Can. J. Math. 19, p. 1110.

Arnett W. D. and Bowers R. L. (1977). 'A Microscopic Interpretation of Neutron Star Structure', Astrophys. J. Suppl. 33, p. 415.

Ashtekar A. (1986). 'New Variables for Classical and Quantum Gravity', Phys. Rev. Lett. 57, p. 2244.

Ashtekar A. (1991). 'Old Problems in the Light of New Variables', in 'Conceptual Problems of Quantum Gravity' (ed. A. Ashtekar and J. Stachel), Birkhauser, Boston.

Ashtekar A. and Hansen R. O. (1978). 'A Unified Treatment of Null and Spatial Infinity in General Relativity. I. Universal Structure, Asymptotic Symmetries, and Conserved quantities at Spatial Infinity', J. Math. Phys. 19, p. 1542.

Ashtekar A. and Stachel J. (ed.) (1991). 'Conceptual Problems of Quantum Gravity' , Birkhauser, Boston.

Avez A. (1963). 'Essais de geometrie Riemanniene hyperbolique globale. Applications a la Relativite Generale', Ann. Inst. Fourier (Grenoble), 132, p. 105.

Bardeen J. M., Carter B. and Hawking S. W. (1973). 'The Four Laws of Black Hole Mechanics', Commun. Math. Phys. 31, p. 161.

Barrabes C., Israel W. and Letelier P. S. (1991). 'Analytic Models of Nonspherical Collapse, Cosmic Censorship and the Hoop Conjecture', Phys. Lett. 160A, p. 41.

Barrow J. D. (1981). 'Cosmological Constraints on Elementary Particle Masses', Phys. Lett. 107B, p. 3518.

Barrow J. D. and Turner M. (1982). 'The Inflationary Universe–Birth, Death and Transfiguration', Nature, 298, p. 801.

Beem J. (1976). 'Conformal Changes and Geodesic Completeness' Commun. Math. Phys. 49, p. 179.

Beem J. and Ehrlich P. (1981). 'Global Lorentzian Geometry', Dekker, New York.

Bekenstein J. D. (1973). 'Black Holes and Entropy', Phys. Rev. D7, p. 2333.

Belinskii V.A., Lipschitz E.M. and Khalatninikov I.M. (1972). 'Construction of a Generalized Cosmological Solution of the Einstein Equations with a Time Singularity', Sov. Phys. JETP, 35, p. 383.

Berezinsky V. S. (1991). Phys. Lett. 261B, p. 71.

Bicknell G. and Henriksen R.N. (1978). 'Self-Similar Growth of Primordial Black Holes. I Stiff Equation of State', Astrophys. J. , 219, p. 1043.

Birkhoff G. D. (1923). 'Relativity and Modern Physics', Harvard University Press, Cambridge, MA.

Bishop R. L. and Critendon R. J. (1964). 'Geometry of Manifolds', Academic Press, New York.

Blandford R. D. (1987). 'Astrophysical Black holes', in 'Three Hundred Years of Gravitation', (ed. S. W. Hawking and W. Israel), Cambridge University Press, Cambridge.

Blandford R. D. and Thorne K. (1979). 'Black Hole Astrophysics', in 'General Relativity—An Einstein Centenary Survey' (ed. S. W. Hawking and W. Israel), Cambridge University Press, Cambridge.

Blumenthal G. R., Faber S. M., Primack J. R., and Rees M. J. (1984). 'Formation of Galaxies and Large Scale Structure with Cold Dark Matter', Nature, 311, p. 517.

Bondi H. (1948). 'Spherically Symmetric Models in General Relativity', Mon. Not. Astron. Soc. 107, p. 400.

Bondi H. and Gold T. (1948). 'The Steady State theory of the Expanding Universe', Mon. Not. Roy. Astro. Soc. 108, p. 252.

Bondi H., van der Burgh M. G. J. and Metzner R. (1962). 'Gravitational Waves in General Relativity VII. Waves from Axi-symmetric Isolated Systems', Proc. Roy. Soc. Lond. A269, p. 21.

Borde A. (1987). 'Geodesic Focusing, Energy Conditions and Singularities', Class. Quantum Grav. 4, p. 343.

Boyer R. H. and Lindquist R. W. (1967). 'Maximal Analytic Extension of Kerr Metric', J. Math. Phys. 8, p. 265.

Brans C. and Dicke R. H. (1961). 'Mach's Principle and a Relativistic Theory of Gravitation', Phys. Rev. 124, p. 925

Brill D. (1975). 'Quantum Cosmology', in 'Quantum Theory and Structure of Time and Space' (ed. L. Castel, M. Drieschner, and C. F. von Weizsacker), Carl Hansen Verlag, Munich.

Budic R. and Sachs R. K. (1974). 'Causal Boundaries for General Relativistic Space-times', J. Math. Phys. 15, p. 1302.

Cahill M. E. and Taub A. H. (1971). 'Spherically Symmetric Similarity Solutions of the Einstein Field Equations for a Perfect Fluid', Commun. Math. Phys. 21, p. 1.

Carr B. and Hawking S. W. (1974). 'Black Holes in the Early Universe', Mon. Not. Roy. Astro. Soc., 168, p. 399.

Carter B. (1968). 'Global Structure of the Kerr Family of Gravitational Field', Phys. Rev. 174, p. 1559.

Carter B. (1971). 'Causal Structure in Space-time', Gen. Relat. and Grav. 1, p. 349.

Carter B. (1979). 'The General Theory of the Mechanical, Electromagnetic and Thermodynamic Properties of Black Holes' in 'General Relativity—An Einstein Centenary Survey' (ed. S. W. Hawking and W. Israel), Cambridge University Press, Cambridge.

Chandrasekhar S. (1931). 'The Maximum Mass of Ideal White Dwarfs', Astrophys. J. 74, p. 81.

Chandrasekhar S. (1934). 'Stellar Configurations with Degenerate Cores', Observatory 57, p. 373.

Chandrasekhar S. (1983). 'Mathematical Theory of Black Holes', Clarendon Press, Oxford.

Chandrasekhar S. and Hartle J. (1983). 'On Crossing the Cauchy Horizon of a Reissner-Nordström Black hole', Proc. Roy. Soc. Lond. A384, p. 301.

Christodoulou D. (1970). 'Reversible and Irreversible Transformations in Black Hole Physics', Phys. Rev. Lett. 25, p. 1596.

Christodoulou D. (1984). 'Violation of Cosmic Censorship in the Gravitational Collapse of a Dust Cloud', Commun Math. Phys. 93, p. 171.

Christodoulou D. (1986). 'Global Existence of Generalized Solutions of the Spherically Symmetric Einstein-Scalar Equations in the Large', Commun. Math. Phys. 106, p. 587.

Clarke C. J. S. (1971). 'On the Geodesic Completeness of Causal Space-times', Proc. Camb. Phil. Soc. 69, p. 319.

Clarke C. J. S. (1975a). 'Singularities in Globally Hyperbolic Space-times', Commun. Math. Phys. 41, p. 65.

Clarke C. J. S. (1975b). 'The Classification of Singularities', Gen. Relat. Grav. 1, p. 35.

Clarke C. J. S. (1986). 'Singularities: Global and Local Aspects', in 'Topological Properties and Global Structure of Space-time' (ed. P. G. Bergmann and V. de Sabbata), Plenum Press, New York.

Clarke C. J. S. and de Felice F. (1982). 'Globally Non-causal Space-times', J. Phys. A15, p. 2415.

Clarke C. J. S. and de Felice F. (1984). 'Globally Non-causal Space-times II, Naked Singularities and Curvature Conditions', Gen. Relat. Grav. 16, p. 139.

Clarke C. J. S. and Krolak A. (1986). 'Conditions for the Occurrence of Strong Curvature Singularities', J. Geo. Phys. 2, p. 127.

Clarke C. J. S. and Joshi P. S. (1988). 'On Reflecting Space-times', Class. Quantum Grav. 5, p. 19.

Cline and Raby (1991). 'Gravitino Induced Baryogenesis: A Problem made a virtue', Phys. Rev. D, 43, p. 1781.

Cowan J. J. , Thielmann F.-K. and Truran J. W. (1987). 'Nuclear Chronometers from the r-processes and the Age of the Galaxy', Astrophys. J. 323, p. 543.

Cowsik R. and MacClelland J. (1972). 'An Upper Limit on the Neutrino Rest Mass', Phys. Rev. Lett., 29, p. 669.

Cruz de la V., Chase J.E. and Israel W. (1970). 'Gravitational Collapse with Symmetries', Phys. Rev. Lett. 24, p. 423.

Datta B. and Joshi P. S. (1986). 'Lower Bounds on Axion Rest Mass in a General Cosmological Scenario', Bull. Astro. Soc. India, 14, p. 156.

Davies T. and James E. (1966). 'Non-linear Differential Equations, ' Addison Wesley, New York.

Deser S., van Nieuwenhuizen N. and Boulware D. (1975). 'Uniqueness and Non-renormalizability of Quantum Gravitation', in 'General Relativity and Gravitation' (ed. G. Shaviv and N. Rosen), Wiley Pub. New York.

Deutsch D. and Candelas P. (1980). 'Boundary Effects in Quantum Field Theory', Phys. Rev. D, 20, p. 3063.

Dewitt B. (1967). 'Quantum Theory of Gravity: The Menifestly Covariant Theory', Phys. Rev., 162, p. 1195.

Dewitt B. (1979). 'Quantum Gravity: the New Synthesis' in 'General Relativity-An Einstein Centenary Survey' (ed. S. W. Hawking and W. Israel), Cambridge University. Press, Cambridge.

Dhurandhar S. V. and Joshi P. S. (1981). 'Causality Conditions and the Lengths of non-spacelike Curves', Gen. Relat. and Grav. 13, p. 913.

Dine M. and Fischler W. (1983). 'The Not-so-harmless Axion', Phys. Lett. 120B, p. 137.

Dixon W. G. (1970). 'Dynamics of Extended Bodies in General Relativity: Momentum and Angular Momentum', Proc. Roy. Soc. Lond., A314, p. 499.

Dolgov A. D. and Zeldovich Ya. B. (1981). 'Cosmology and Elementary Particles', Rev. Mod. Phys. 53, p. 1.

Doroshkevich A. G., Zeldovich Ya. B., and Novikov I. (1966). 'Gravitational Collapse of non-symmetric non-rotating Masses', Sov. Phys. JETP, 22, p. 122.

Dray T. and Joshi P. S. (1990). 'Gluing Reissner-Nordstrom Space-times along Charged Shells of Matter', Class. Quantum Grav. 7, p. 41.

Dwivedi I. H. and Joshi P. S. (1989). 'On the Nature of Naked Singularities in Vaidya space-times', Class. Quantum Grav. 6, p. 1599.

Dwivedi I. H. and Joshi P. S. (1991). 'On the Nature of Naked Singularities in Vaidya Space-times II', Class. Quantum Grav. 8, p. 1339.

Dwivedi I. H. and Joshi P. S. (1992). 'Cosmic censorship Violation in Non-self-similar Tolman–Bondi Models', Class. Quantum Grav. 9, p. L69.

Eardley D. (1974). 'Self-Similar Space-times: Geometry and Dynamics', Commun. Math. Phys. 37, p. 287.

Eardley D. M. (1987). 'Naked Singularities in Spherical Gravitational Collapse', in 'Gravitation in Astrophysics', (ed. B. Carter and J. B. Hartle), Plenum, New York.

Eardly D. M. and Smarr L. (1979). 'Gravitational Collapse of Dust Spheres: Time Coordinate Conditions', Phys., Rev. D19 p. 2239.

Ellis G. F. R. and King A. R. (1974). 'Was the Big bang a Whimper?' Commun. Math. Phys., 38, p. 119.

Ellis G. F. R. and Schmidt B. (1977). 'Singular Space-times', Gen. Relat. and Grav., 8, p. 915.

Exton A., Newman E. T., and Penrose R. (1969). 'Conserved Quantities in Einstein-Maxwell Theory', J. Math. Phys., 10, p. 1566.

Feynman R. P. and Hibbs (1965). 'Quantum Mechanics and Path Integrals', MacGraw-Hill, New York.

Finkelstein D. (1958). 'Past-future Asymmetry of the Gravitational Field of a point particle', Phys. Rev., 110, p. 965.

Fock V. (1939). 'Sur le Mouvement des Masses Finies d'apres la Theorie Gravitation Einsteinienne', J. Phys. U.S.S.R., 1, p. 81.

Ford L. H. and Parker L. (1978). 'Creation of Particles by Singularities in Asymptotically Flat Space-times', Phys. Rev. D, 17, p. 1485.

Friedmann A. (1922). 'Uber die Krummung des Raumes', Z. Phys. 10, p. 377.

Friedman J. L. (1991). 'Space-time Topology and Quantum Gravity', in 'Conceptual Problems of Quantum Gravity' (ed. A. Ashtekar and J. Stachel), Birkhauser, Boston.

Friedman J., Morris M.S., Novikov I.D., Echeverria F., Klinkhammer G., and Thorne K.S., 'Cauchy Problem in Space-times with Closed Timelike Curves' , Phys. Rev. D42, p. 1915.

Galloway G. (1983). 'Causality Violation in Spatially Closed Space-times', Gen. Relat. and Grav. 15, p. 165.

Geller M. J. and Huchra J. P. (1989). 'Mapping the Universe', Science, 246, p. 897.

Geroch R. (1967). 'Topology in General Relativity', Jour. Math. Phys. 8, p. 782.

Geroch R. (1968a). 'What is a Singularity in General Relativity?' Ann. Phys., 48, p. 526.

Geroch R. (1968b). 'Spinor Structure of Space-times in General Relativity I' , J. Math. Phys., 9, p. 1739.

Geroch R. (1970a). 'Domain of Dependence', Jour. Math. Phys. 11, p. 437.

Geroch R. (1970b). 'Singularities', in 'Relativity' (ed. S. Fickler, M. Carmeli and L. Witten), Plenum Press, New York.

Geroch R. (1971). 'Space-time Structure from a Global View Point', in 'General Relativity and Cosmology', Proceedings of the International School in Physics 'Enrico Fermi' (ed. R. K. Sachs), Academic Press, New York.

Geroch R. and Horowitz G. (1979). 'Global Structure of Space-times', in 'General Relativity—an Einstein Centenary Survey' (ed. S. W. Hawking and W. Israel), Cambridge University Press, Cambridge.

Geroch R., Kronheimer E. and Penrose R. (1972). 'Ideal Points in Space-time', Proc. R. Soc. Lond. A327, p. 545.

Gerstein S. S. and Zeldovich Ya. B. (1966). 'Rest Mass of Muonic Neutrino and Cosmology', Sov. Phys. JETP Lett. 4, p. 120

Gibbons G. (1972). 'The Time Symmetric Initial Value Problem for Black Holes', Commun. Math. Phys., 27, p. 87.

Gibbons G., Hawking S. W., Horowitz G., and Perry M. J. (1983). 'Positive Mass Theorems for Black Holes', Commun. Math. Phys., 88, p. 295.

Godel K. (1949). 'An Example of a New type of Cosmological Solution of Einstein's field Equations of Gravitation', Rev. Mod. Phys., 21, p. 447.

Gotay M. J. and Demaret J. (1983). 'Quantum Cosmological Singularities', Phys. Rev. D28, p. 2402.

Gowdy R. H. (1977). 'Instantaneous Cauchy Surfaces, Topology Change and Exploding Black Holes', Jour. Math. Phys. 18, p. 1798.

Graves J. C. and Brill D. (1960). 'Oscillatory character of Reissner-Nordström Metric for an Ideal Charged Worm Hole', Phys. Rev. 120, p. 1507.

Green M., Schwarz J. and Witten E. (1987). 'Superstring Theory', Cambridge University Press, Cambridge.

Grillo G. (1991). 'On a Class of Naked Strong-Curvature Singularities', Class. Quantum Grav. 8, p. 739.

Guth A. (1981). 'Inflationary Universe: A Possible Solution to the Horizon and Flatness Problems', Phys. Rev. D23, p. 347.

Hagerdorn R. (1968). 'Hadronic Matter near the Boiling Point', Nuovo Cimento 56A, p. 1027.

Hartle J. (1978). 'Relativistic Stars, Gravitational Collapse, and Black Holes', in 'Relativity, Astrophysics and Cosmology' (ed. W. Israel), Reidel, Dordrecht.

Hartle J. (1986). 'Quantum Cosmology', in 'High Energy Physics 1985' Proceedings of Yale Theoretical Advanced Study Institute, (ed. M. J. Bowick and F. Gursey), World Scientific, Singapore.

Harrison B. K., Thorne K. S., Wakano M., and Wheeler J., (1965). 'Gravitation Theory and Gravitational Collapse', The University of Chicago Press, Chicago.

Hawking S. W. (1967). 'The Occurrence of Singularities in Cosmology III: Causality and Singularities', Proc. Roy. Soc. Lond., A300, p. 187.

Hawking S. W. (1971). 'Gravitational Radiation from Colliding Black Holes', Phys. Rev. Lett., 26, p. 1344.

Hawking S. W. (1975). 'Particle Creation by Black Holes' Commun. Math. Phys., 43, p. 199.

Hawking S. W. (1992). 'Evaporation of Two-dimensional Black Holes', Phys. Rev. Lett., 69, p. 406.

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

Hawking S. W. and Hartle J. (1972). 'Energy and Angular Momentum Flow into a Black Hole', Commun. Math. Phys., 27, p. 283.

Hawking S. W. and Israel W. (ed.) (1979). 'General Relativity- An–Einstein Centenary Survey', Cambridge Univ. Press, Cambridge.

Hawking S. W. and Israel W. (1979). 'An Introductory Survey' in 'General Relativity–An Einstein Centenary Survey' (ed. S. W. Hawking and W. Israel), Cambridge University. Press, Cambridge.

Hawking S. W. and Penrose R. (1970). 'The Singularities of Gravitational Collapse and Cosmology', Proc. R. Soc. Lond. A314, p. 529.

Hawking S. W. and Rosek R. (ed.)(1981). 'Superspace and Supergravity', Cambridge University Press, Cambridge.

Hawking S. W. and Sachs R. K. (1974). 'Causally Continuous Space-times', Commun. Math. Phys. 35, p. 287.

Hawking S. W., King A. R. and McCarthy P. J. (1976). 'A New Topology for Curved Space-time which Incorporates the Causal, Differential and Conformal Structures', J. Math. Phys. 17, p. 174.

Hellaby J. and Lake K. (1985). 'Shell Crossings and the Tolman Model', Astrophys. J. 290, p. 381.

Hicks N. J. (1965). 'Notes on Differential Geometry', Van Nostrand, Princeton, New Jersey.

Hille E. (1969). 'Lectures on Ordinary Differential Equations', Addison-Wesley Pub. New York.

Hiscock W. A., Williams L. G., and Eardley D. M. (1982). 'Creation of Particles by Shell-focusing Singularities', Phys. Rev. D 26, p. 751.

Hollier G. (1986). 'Papapetrou's Naked Singularity is a Strong Curvature Singularity', Class. Quantum Grav. 3, L111.

Hoyle F. (1948). 'A New Model for the Expanding Universe', Mon. Not. Roy. Astro. Soc., 108, p. 372.

Hoyle F. and Narlikar J. V. (1964). 'A new Theory of Gravitation', Proc. Roy. Soc. Lond., A282, p. 191.

Isham C. J., Penrose R. and Sciama D. W. (1981). 'Quantum Gravity 2: A Second Oxford Symposium', Clarendon Press, Oxford.

Ishikawa H. (1979). 'Properties of the Set of Reflecting Points', Gen. Relat. Grav. 10, p. 31.

Israel W. (1984). 'Does a Cosmic Censor Exist?', Found. Phys. 14, p. 1049.

Israel W. (1986a). 'Must Non-spherical Collapse Produce Black Holes ? A Gravitational Confinement Theorem', Phys. Rev. Lett. 56, p. 789.

Israel W. (1986b). 'The Formation of Black Holes in non-spherical Collapse and Cosmic Censorship', Can. J. Phys. 64, p. 120.

Jang P. S. and Wald R. (1977). 'The Positive Energy Conjecture and the Cosmic Censor Hypothesis', J. Math. Phys., 18, p. 41.

Joshi P. S. (1980). 'Space-time Singularities and Microwave Background Radiation', Pramana, 15, p. 225.

Joshi P. S. (1981). 'On Higher Order Causality Violations', Phys. Lett. 85A, p. 319.

Joshi P. S. (1986). 'General Upperlimits to the Age of the Universe', Bull. Astro. Soc. of India, 14, p. 1.

Joshi P. S. and Chitre S. M. (1981a). 'Neutrinos of non-zero Mass in Friedmann Universes', Nature, 293, p. 679.

Joshi P. S. and Chitre S. M. (1981b). 'Upper Bounds on Neutrino Masses from the Large Scale Structure of Space-time', Phys. Lett., 85A, p. 131.

Joshi P. S. and Dwivedi I. H. (1991). 'Strengths of Naked Singularities in Radiation Collapse with non-linear Mass functions', J. Math. Phys. 32, p. 2167.

Joshi P. S. and Dwivedi I. H. (1992a). 'Strong Curvature Naked Singularities in Non-self-similar Gravitational Collapse', Gen. Relat. Grav. 24, p. 129.

Joshi P. S. and Dwivedi I. H. (1992b). 'Naked Singularities in Non-self-similar Gravitational Collapse of Radiation Shells', Phys. Rev. D 45, p. 2147.

Joshi P. S. and Dwivedi I. H. (1992c). 'The Structure of Naked Singularity in Self-similar Gravitational Collapse', Commun. Math. Phys. 146, p. 333.

Joshi P. S. and Joshi S. S. (1987). 'Quantum Effects in a Homogeneous Dust Cloud Collapse', Gen. Relat. Grav. 19, p. 1033.

Joshi P. S. and Joshi S. S. (1988). 'Quantum Effects near a Black Hole Singularity', Class. Quantum Grav. 5, L191.

Joshi P. S. and Narlikar J. V. (1982). 'Black Hole Physics in Globally Hyperbolic Space-times', Pramana J. of Phys. 18, p. 385.

Joshi P. S. and Narlikar J. V. (1986). 'Conformal Quantization and Space-time Singularity', Int. Jour. of Mod. Physics A, 1, p. 243.

Joshi P. S. and Newman E. T. (1984). 'Light Cone Cuts for Charged Kerr Geometries', Gen. Relat. and Grav., 16, p. 1157.

Joshi P. S. and Saraykar R. V. (1987). 'Cosmic Censorship and Topology Change in General Relativity', Phys. Lett. 120A, p. 111.

Joshi P. S , Kozameh C. N. and Newman E. T. (1983). 'Light Cone Cuts of Null Infinity in Schwarzschild Geometry', Jour. Math. Phys. 24, p. 2490.

Joshi P. S. , Padmanabhan T. and Chitre S. M. (1987). 'Bounds on Vacuum Energy Density in a General Cosmological Scenario', Phys. Lett. 120A, p. 115.

Kahana S., Baron E. and Cooperstein J. (1984). 'Successful Supernovae, the Anatomy of Shocks: Neutrino Emission and the Adiabatic Index' in 'Problem Collapse and Numerical Relativity' (ed. D. Bancel and M. Signore), Reidel, Dordrecht.

Kerr R. P. (1963). 'Gravitational Field of a Spinning Particle as an Example of Algebrically Special Metrics', Phys. Rev. Lett. 11, p. 237.

Khlopov M. Yu and Linde A. D. (1984). 'Is it Easy to Save the Gravitino ?', Phys. Lett. 138B, p. 265.

King A. R. (1975). 'Instability of Intermidiate Singularities in General Relativity', Phys. Rev. D, 11, p. 763.

Kinnersley W. (1969). 'Type D Vacuum Metrics', J. Math. Phys., 10, p. 1195.

Kolb E. and Turner M. (1988). 'The Early Universe' (Reprints), Addisson-Wesley, Reading Massachusetts.

Kozameh C. N. and Newman E. T. (1986). 'A Non-local Approach to the Vacuum Maxwell, Yang-Mills, and Einstein Equations', in 'Topological Properties and Global Structure of Space-time' (ed. P. G. Bergmann and V. de Sabbata), Plenum Press, New York.

Kriele M. (1990). 'Causality Violation and Singularities', Gen. Relat. Grav. 22, p. 619.

Królak A. (1983). 'A Proof of the Cosmic Censorship Hypothesis', Gen. Relat. Grav. 15, p. 99.

Kruskal M. D. (1960). 'Maximal Extension of Schwarzschild Metric', Phys. Rev. 119, p. 1743.

Kuchar K. (1981). 'Canonical Methods of Quantization' in 'Quantum Gravity 2: An Oxford Symposium' (ed. C. J. Isham, R. Penrose, and D. W. Sciama), Clarendon Press, Oxford.

Kuroda Y. (1984). 'Naked Singularities in Vaidya Space-times', Prog. Theor. Phys. 72, p. 63.

Lake K. (1986). 'Smooth (C^2) Formation of a Persistent Naked Singularity', Phys. Lett. A116, p. 17.

Lake K. (1991). 'Naked Singularities in Gravitational Collapse which is not Self-similar', Phys. Rev. D 43, p. 1416.

Lee C. W. (1983). 'The Topology of Geodesically Complete Space-times', Gen. Relat. Grav. 15, p. 21.

Leibovitz C. and Israel W. (1970). 'Maximum Efficiency of Energy Release in Spherical Collapse', Phys. Rev. , D1, p. 3226.

Lemos J. P. S. (1992). 'Naked Singularities: Gravitationally Collapsing Configurations of Dust or Radiation in Spherical Symmetry, a Unified Treatment', Phys. Rev. Lett. 68, p. 1447.

Leray J. (1952). 'Hyperbolic Differential Equations', Institute for Advanced Study, Princeton, Preprint.

Lester J. A. (1984). 'The Causal Automorphism of de Sitter and Einstein Cylinder Space-times', J. Math. Phys. 25, p. 113.

Lin C. C., Mestel L. and Shu F. H. (1965). 'The Gravitational Collapse of a Uniform Spheroid', Astrophys. J., 142, p. 1431.

Linde A. (1991). 'Inflationary Cosmology', in 'Physical Cosmology', ed. A. Blanchard *et al.* Editions Frontieres, France.

Lindquist R. W. , Schwartz R. A. and Misner C. W. (1965). 'Vaidya's Radiating Schwarzschild Metric', Phys. Rev. 137, p. B 1364.

Lovelock D. (1972). 'The Four-Dimensionality of Space and the Einstein Tensor', J. Math. Phys., 13, p. 874.

Ludvigsen M. and Vickers J. A. G. (1983). 'A Simple Proof of the Positivity of Bondi Mass', J. Phys. A, 15, L67.

MacCallum M. A. H. (1979). 'Anisotropic and Inhomogeneous Relativistic Cosmologies', in 'General Relativity—An Einstein Centenary Survey' (ed. S. W. Hawking and W. Israel), Cambridge University. Press, Cambridge.

McNamara J. M. (1978). 'Instability of Black Hole Inner Horizons', Proc. Roy. Soc. Lond., A358, p. 499.

Malament D. (1977). 'The Class of Continuous Timelike Curves determines the Topology of Space-time', Jour. Math. Phys. 18, p. 1399.

May M. M. and White R. H. (1966). 'Hydrodynemicak Calculations of General Relativistic Collapse', Phys. Rev. 141, p. 1232.

Messiah A. (1973). 'Quantum Mechanics' Vol. I, North Holland, Amsterdam.

Misner C. (1963). 'The Flatter Regions of Newman, Unti and Tamburino's Generalrized Schwarzschild space', J. Math. Phys., 4, p. 924.

Misner C. (1967). 'Taub-NUT space as a Counterexample to Almost Anything', in 'Relativity Theory and Astrophysics I: Relativity and Cosmology', (ed. J. Ehlers) Lectures in Applied Mathematics, Vol. 8, American Methematical Society, Providence.

Misner C. and Sharp D. H. (1964). 'Relativistic Equations for Adiabatic, Spherically Symmetric Gravitational Collapse', Phys. Rev., 136, B571.

Misner C. and Taub A.H. (1969). 'A Singularity Free Empty Universe', Sov. Phys. J.E.T.P. 28, p. 122.

Misner C., Thorne, K. and Wheeler J. (1973). 'Gravitation', Freeman, San Francisco.

Morris M. S. , Thorne K. S. and Yurtsever U. (1988). 'Wormholes, Time Machines, and the Weak Energy Condition', Phys. Rev. Lett. 61, p. 1446.

Narlikar J. V. (1981). 'Quantum Conformal Fluctuations near the Classical Spacetime Singularity', Found. Phys., 11, p. 473.

Narlikar J. V. (1983). 'Introduction to Cosmology', Jones and Bartlett, Boston.

Narlikar J. V. (1984). 'The Vanishing Likelihood of Space-time Singularity in Quantum Conformal Cosmology', Found. Phys., 14, p. 443.

Narlikar J. V. and Padmanabhan T. (1983). 'Quantum Cosmology via Path Integrals', Phys. Rep. 100, p. 153.

Nemytskii V. V. and Stepanov V. V. (1960). 'Qualitative Theory of Differential Equations', Princeton University Press, Princeton, N.J.

Newman R. P. A. C. (1984). 'Cosmic Censorship and Conformal Transformations', Gen. Relat. Grav. 16, p. 943.

Newman R. P. A. C. (1986a). 'Strengths of Naked Singularities in Tolman-Bondi Space-times', Class. Quantum Grav. 3, p. 527.

Newman R. P. A. C. (1986b). 'Cosmic Censorship and Strengths of Singularities' in 'Topological Properties and Global Structure of Space-time' (ed. P. G. Bergmann and V. de Sabbata), Plenum Press, New York.

Newman R. P. A. C. (1989). 'Black Holes without Singularities', Gen. Relat. and Grav., 21, p. 981.

Newman R. P. A. C. and Clarke C. J. S. (1987). 'An $\mathbb{R}^4$ space-time with a Cauchy surface which is not $\mathbb{R}^3$', Class. Quantum Grav. 4, p. 53.

Newman R. P. A. C. and Joshi P. S. (1988). 'Constraints on the Structure of Naked Singularities in Classical General Relativity', Ann. Phys. 182, p. 112.

Newman E. T. and Penrose R. (1966). 'Note on Bondi-Metzner-Sachs Group', J. Math. Phys. 7, p. 863.

Newman E. T. and Tod K. P. (1980). 'Asymptotically Flat Space-times', in 'General Relativity and Gravitation' Vol. 2, (ed. A. Held), Plenum, New York.

Newman E. T. , Tamburino L. and Unti T. J. (1963). 'Empty Space Generalization of the Schwarzschild Metric', J. Math. Phys., 4, p. 915.

Newman E. T., Couch E., Chinnapared K., Exton A., Prakash A., and Torrence R. (1965). 'Metric of a Rotating Charged Mass', J. Math. Phys. 6, p. 918.

Oppenheimer J. R. and Snyder H. (1939). 'On Continued Gravitational Contraction', Phys. Rev. 56, p. 455.

Ori A. (1992). 'Inner Structure of a Charged Black Hole: An Exact Mass-Inflation Solution', Phys. Rev. Lett. 67, p. 789.

Ori A. and Piran T. (1987). 'Naked Singularity in self-similar Spherical Gravitational Collapse', Phys. Rev. Lett. 59, p. 2137.

Ori A. and Piran T. (1990). 'Naked Singularities and other features of Self-similar General Relativistic Gravitational Collapse', Phys. Rev. D42, p. 1068.

Pagels H. and Primack J. R. (1982). 'Supersymmetry, Cosmology, and New Physics at Teraelectronvolt Energies', Phys. Rev. Lett. 48, p. 223.

Papapetrou A. (1985). 'Formation of a Singularity and Causality', in 'A Random Walk in Relativity and Cosmology' (ed. N. Dadhich, J. Krishna Rao, J. V. Narlikar, C. V. Vishveshwara), Wiley Eastern, New Delhi.

Papapetrou A. and Hamoui A. (1967). 'Surfaces Caustiques degenerees dans la solution de Tolman. La Singularite physique en Relativite Generale', Ann. Inst. H Poincare AVI p. 343.

Peccei R. and Quinn H. (1977). 'CP Conservation in the Presence of Pseudo Particles', Phys. Rev. Lett. 38, p. 1440.

Penrose R. (1965). 'Zero Rest Mass Fields Including Gravitation: Asymptotic Behavior', Proc. Roy. Soc. Lond. A284, p. 159.

Penrose R. (1968). 'Structure of Space-time' in 'Battelle Rencontres', (ed. C. Dewitt and J. A. Wheeler), Benjamin, New York.

Penrose R. (1969). 'Gravitational Collapse : The Role of General Relativity', Riv. Nuovo Cimento Soc. Ital. Fis. 1, p. 252.

Penrose R. (1972). 'Techniques of Differential Topology in Relativity' A.M.S. Colloquium Publications, SIAM, Philadelphia.

Penrose R. (1974a). 'Gravitational Collapse' in 'Gravitational Radiation and Gravitational Collapse' (IAU Symposium No. 64) ed. C. DeWitt-Morettee, Reidel, Dordrecht.

Penrose R. (1974b). 'Singularities in Cosmology' in 'Confrontation of Cosmological Theories with Observational Data' ed. M. S. Longair, Reidel, Dordrecht.

Penrose R. (1979). 'Singularities and Time asymmetry', in 'General Relativity-an Einstein Centenary Survey' (ed. S. W. Hawking and W. Israel), Cambridge University, Press, Cambridge.

Penrose R. (1982). 'Quasi-local Mass and Angular Momentum in General Relativity', Proc. Roy. Soc. Lond. A381, p. 53.

Perko L. (1991). 'Differential Equations and Dynamical Systems', Springer-Verlag, New York.

Preskill J. , Wise M. B. and Wilczek F. (1983). 'Cosmology of Invisible Axion', Phys. Lett. 120B, p. 127.

Press R. H. and Teukolsky S. A. (1973). 'Perturbations of a Rotating Black Hole. Dynamical Stability of the Kerr Metric', Astrophys. J. 185, p. 649.

Price R. H. (1972). 'Non-spherical Perturbations of Relativistic Gravitational Collapse. I. Scalar and Gravitational Perturbations', Phys. Rev. D5, p. 2419.

Primack J. R. (1985). 'Early Universe', in the Proc. of Int. School Phys. 'Enrico Fermi', Verena, Italy.

Racz I. (1987). 'Distinguishing Properties of Causality Conditions', Gen. Relat. Grav. 19, p. 1025.

Rajagopal K. and Lake K. (1987). 'Strengths of Singularities in Vaidya Spacetimes', Phys. Rev. D, 35, p. 1531.

Raychaudhuri A. K. (1955). 'Relativistic Cosmology', Phys. Rev., 98, p. 1123.

Raychaudhri A. K. (1979). 'Theoretical Cosmology', Clarendon Press, Oxford.

Robertson H. P. (1933). 'Relativistic Cosmology' , Rev. Mod. Phys. 5, p. 62.

Roy P., Joshi P. S. and Chitre S. M. (1991). 'Gravitino Mass Bounds in a General Cosmological Scenario', Phys. Lett., 160A, (1991).

Sachs R. (1962). 'Gravitational Waves in General Relativity, VIII. Waves in Asymptotically Flat Space-time', Proc. Roy. Soc. Lond., A270, p. 103.

Sachs R. H. and Wu (1977). 'General Relativity for Mathematicians', Springer-Verlag, New York.

Sandage A. and Tammann (1984). in the First ESO-CERN Symp. on Large Scale Structure of the Universe, Cosmology and Fundamental Physics', (ed. G. Setti and L. van Hove), CERN.

Sandage A. R. and Cacciari C. (1990). 'The Absolute Magnitudes of RR Lyrae Stars and the Age of the Galactic Globular Cluster system', Astrophys. J. 350, 645.

Saunders W., Frenk C. , Rowan-Robinson M. , Efstathiou G. , Lawrence A., Kaiser N., Ellis R., Crawford J., Xia X. Y., and Parry I. (1991). 'The Density Field of the Local Universe', Nature, 349, p. 32.

Schoen R. and Yau S.-T. (1983). 'The Existence of a Black Hole due to Condensation of Matter', Commun. Math. Phys. 90, p. 575.

Sciama D. W. (1990). 'Consistent Neutrino Masses from Cosmology and Solar Physics', Nature, 348, p. 617.

Seifert H. J. (1967). 'Global Connectivity by Timelike Geodesics', Zs. F. Naturfor., 22a, p. 1356.

Seifert H. J. (1971). 'Causal Boundary for Space-times', Gen. Relat. and Grav., 1, p. 247.

Shapiro S. L. and Teukolsky S. A. (1991). 'Formation of Naked Singularities: The Violation of Cosmic Censorship', Phys. Rev. Lett. 66, p. 994.

Simmons G. F. (1963). 'Introduction to Topology and Modern Analysis', McGraw-Hill Kogakusha Ltd., Tokyo.

Simpson M. and Penrose R. (1973). 'Internal Instability in a Reissner- Nordstrom Black Hole', Int. J. Theor. Phys. 7, p. 183.

Symbalisty E. M. D. , Yang J. and Schramm D. (1980). 'Neutrinos and the Age of the Universe', Nature 288, p. 143.

Szalay and Marx (1976). 'Neutrino Rest Mass from Cosmology', Astron. Astrophys. 49, p. 437.

Szekeres P. (1960). 'On the Singularities of a Riemmanian Manifold', Publ. Math. Debrecen, 7, p. 285.

Taub A. H. (1951). 'Empty space-times admitting a three-parameter Group of Motions', Ann. of Math. 53, p. 472.

Thielmann F. K., Metzinger J. and Klapdor H.V. (1983). 'New Actinide Chronometer Production Ratios and the Age of the Galaxy', Astron. Astrophys. 123, p. 162.

Thorne K. S. (1972). 'Non-spherical Gravitational Collapse: A Short Review' in 'Magic without Magic—John Archibald Wheeler', ed. J. Clauder, W. H. Freeman, New York.

Tipler F. (1976). 'Causality Violation in Asymptotically Flat Space-times', Phys. Rev. Lett., 37, p. 879.

Tipler F. (1977a). 'Singularities and Causality Violation', Ann. Phys. 108, p. 1.

Tipler F. (1977b). 'Singularities in Conformally Flat Space-times' Phys. Lett. 64A, p. 8.

Tipler F. (1977c). 'Black Holes in Closed Universes', Nature, 270, p. 500.

Tipler F. (1978). 'General Relativity and Conjugate Ordinary Differential Equations', Jour. Diff. Equations, 30, p. 165.

Tipler F., Clarke C. J. S. and Ellis G. F. R. (1980). 'Singularities and Horizons' in 'General relativity and Gravitation' Vol 2, (ed. A Held) Plenum, New York.

Tolman R. C. (1934). 'Effect of Inhomogeneity on Cosmological Models' Proc. Natl. Acad Sci USA 20, p. 169.

Tricomi F.G. (1961). 'Differential Equations', Blackie and Sons, London.

Vaidya P. C. (1943). 'The External Field of a Radiating Star in General Relativity', Curr. Sci. 12, p. 183.

Vaidya P. C. (1951). 'The Gravitational Field of a Radiating Star', Proc. of the Indian Acad. Sci. A33, p. 264.

Vaidya P. C. (1953). 'Newtonian Time in General Relativity', Nature, 171, p. 260.

Vaucouleurs de G. (1970). 'The Case for a Hierarchical Cosmology', Science, 167, p. 1203.

Verhulst F. (1990). 'Non-linear Differential Equations and Dynamical Systems', Springer-Verlag, Berlin.

Vyas U. D. and Akolia G. M. (1984). 'Chronal Isomorphism', Gen. Relat. Grav. 16, p. 1045.

Vyas U. D. and Akolia G. M. (1986). 'Causally Discontinuous Space-times', Gen. Relat. Grav. 18, p. 309.

Vyas U. D. and Joshi P. S. (1983). 'Causal Functions in General Relativity', Gen. Relat. and Grav., 15, p. 553.

Vyas U. D. and Joshi P. S. (1989). 'Topological Techniques in General Relativity', in 'Geometry and Topology' (ed. G. Rassias and G. Stratopoulos), World Scientific, Singapore.

Wald R. (1984). 'General Relativity', University of Chicago Press, Chicago.

Walker M. and Penrose R. (1970). 'On Quadratic First Integrals of the Geodesic Equations for type [22] space-times', Commun. Math. Phys., 18, p. 265.

Waugh B. and Lake K. (1988). 'Strengths of Shell-focusing Singularities in Marginally Bound Collapsing Self-similar Tolman Space-times', Phys. Rev. D, 38, p. 1315.

Waugh B. and Lake K. (1989). 'Shell-focusing Singularities in Spherically Symmetric Self-similar Space-times', Phys. Rev. D40, p. 2137.

Weinberg S. (1972). 'Gravitation and Cosmology', John Wiley, New York.

Weinberg S. (1978). 'A New Light Boson?', Phys. Rev. Lett. 40, p. 223.

Weinberg S. (1982). 'Cosmological Constraints on the Scale of Supersymmetry Breaking', Phys. Rev. Lett. 48, p. 1303.

Wheeler J. A. (1962). 'Geometrodynamics', Academic Press, New York.

Wheeler J. A. (1964). 'Geometrodynamics and the Issue of Final State', in 'Relativity, Groups and Topology' (ed. C. Dewitt and B. Dewitt), Gordon and Breach, New York.

Wilczek F. (1978). 'Problem of Strong P and T Invariance in the Presence of Instantons' , Phys. Rev. Lett. 40, p. 279.

Wilczek F. (1991). 'Perspecives on Particle Physics and Cosmology', Physica Scripta, T36, p. 281.

Willard S. (1970). 'General Topology', Addison-Wesley, Reading, Massachusetts.

Winter K. (ed.) (1990). 'Proc. Neutrino 1990', Nucl. Phys. B, Special Number.

Woodhouse N. M. J. (1973). 'The Differential and Causal Structures of Spacetime', Jour. Math. Phys., 14, p. 495.

Yodzis P., Seifert H.-J. and Muller zum Hagen H. (1973). 'On the Occurrence of Naked Singularities in General Relativity', Commun. Math. Phys. 34, p. 135.

Yodzis P., Seifert H.-J. and Muller zum Hagen H. (1974). 'On the Occurrence of Naked Singularities in General Relativity II', Commun. Math. Phys., 37, p. 29.

Zeeman E. C. (1964). 'Causality implies Lorentz Group', J. Math. Phys. 5, p. 490.

Zeeman E. C. (1967). 'The Topology of Minkowski Space', Topology, 6, p. 161.

Zeldovich Ya. B. and Novikov I. (1983). 'Relativistic Astrophysics: Stars and Relativity' Vol. 1, The University of Chicago Press, Chicago.

INDEX