Lecture Notes in Physics

The Lecture Notes in Physics

The series Lecture Notes in Physics (LNP), founded in 1969, reports new developments in physics research and teaching – quickly and informally, but with a high quality and the explicit aim to summarize and communicate current knowledge in an accessible way. Books published in this series are conceived as bridging material between advanced graduate textbooks and the forefront of research to serve the following purposes:

• to be a compact and modern up-to-date source of reference on a well-defined topic;
• to serve as an accessible introduction to the field to postgraduate students and nonspecialist researchers from related areas;
• to be a source of advanced teaching material for specialized seminars, courses and schools.

Both monographs and multi-author volumes will be considered for publication. Edited volumes should, however, consist of a very limited number of contributions only. Proceedings will not be considered for LNP.

Volumes published in LNP are disseminated both in print and in electronic formats, the electronic archive is available at springerlink.com. The series content is indexed, abstracted and referenced by many abstracting and information services, bibliographic networks, subscription agencies, library networks, and consortia.

Proposals should be sent to a member of the Editorial Board, or directly to the managing editor at Springer:

Dr. Christian Caron
Springer Heidelberg
Physics Editorial Department I
Tiergartenstrasse 17
69121 Heidelberg/Germany
christian.caron@springer-sbm.com

Jörg Frauendiener Domenico J.W. Giulini
Volker Perlick (Eds.)

Analytical and Numerical Approaches to Mathematical Relativity

With a Foreword by Roger Penrose

Springer

Editors

Jörg Frauendiener
Institut für Theoretische Astrophysik
Universität Tübingen
Auf der Morgenstelle 10
72076 Tübingen, Germany
E-mail: joergf@tat.physik.uni-tuebingen.de

Volker Perlick
Institut für Theoretische Physik
TU Berlin
Hardenbergstrasse 36
10623 Berlin
E-mail: vper0433@itp.physik.tu-berlin.de

Domenico J.W. Giulini
Fakultät für Physik und Mathematik
Universität Freiburg
Hermann-Herder-Str. 3
79104 Freiburg, Germany
E-mail: giulini@physik.uni-freiburg.de

J. Frauendiener et al., *Analytical and Numerical Approaches to Mathematical Relativity*, Lect. Notes Phys. 692 (Springer, Berlin Heidelberg 2006), DOI 10.1007/b11550259

Library of Congress Control Number: 2005937899

ISSN 0075-8450
ISBN-10 3-540-31027-4 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-31027-3 Springer Berlin Heidelberg New York

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

Printed in The Netherlands

Typesetting: by the authors and TechBooks using a Springer LaTeX macro package

Printed on acid-free paper SPIN: 11550259 54/TechBooks 5 4 3 2 1 0

Foreword

The general theory of relativity, as formulated by Albert Einstein in 1915, provided an astoundingly original perspective on the physical nature of gravitation, showing that it could be understood as a feature of a curvature in the four-dimensional continuum of space-time. Now, some 90 years later, this extraordinary theory stands in superb agreement with observation, providing a profound accord between the theory and the actual physical behavior of astronomical bodies, which sometimes attains a phenomenal precision (in one case to about one part in one hundred million million, where several different non-Newtonian effects, including the emission of gravitational waves, are convincingly confirmed). Einstein's tentative introduction, in 1917, of an additional term in his equations, specified by a "cosmological constant", appears now to be observationally demanded, and with this term included, there is no discrepancy known between Einstein's theory and classical dynamical behavior, from meteors to matter distributions at the largest cosmological scales. One of Einstein's famous theoretical predictions that light is bent in a gravitational field (which had been only roughly confirmed by Eddington's solar eclipse measurements at the Island of Principe in 1919, but which is now very well established) has become an important tool in observational cosmology, where gravitational lensing now provides a unique and direct means of measuring the mass of very distant objects.

But long before general relativity and cosmology had acquired this impressive observational status, these areas had provided a prolific source of mathematical inspiration, particularly in differential geometry and the theory of partial differential equations (where sometimes this had been applied to situations in which the number of space-time dimensions differs from the four of direct application to our observed space-time continuum). As we see from several of the articles in this book, there is still much activity in all these mathematical areas, in addition to other areas which have acquired importance more recently. Most particularly, the interest in black holes, with their horizons, their singularities, and their various other remarkable properties, both theoretical and in relation to observed highly dramatic astronomical phenomena, has also stimulated much important research. Some have interesting mathematical implications, involving particular types of mathematical argumentations, such as the involvement of differential topology and

the study of families of geodesics, and some having relevance to deep foundational issues relating to quantum theory and thermodynamics. We find a good representation of these discussions here. Some distinct progress in the study of asymptotically flat space-times is also reported here, which greatly clarifies the issue of what can and cannot be achieved using the method of conformal compactification.

In addition to (and sometimes in conjunction with) such purely mathematical investigation, there is a large and important body of technique that has grown up, which has been made possible by the astonishing development of electronic computer technology. Enormous strides in the computer simulation of astrophysical processes have been made in recent years, and this has now become an indispensable tool in the study of gravitational dynamics, in accordance with Einstein's general relativity (such as with the study of black-hole collision that will form an essential part of the analysis of the signals that are hoped to be detected, before too long, by the new generations of gravitational wave detectors). Significant issues of numerical analysis inevitably arise in conjunction with the actual computational procedures, and issues of this nature are also well represented in the accounts presented here.

It will be seen from these articles that research into general relativity is a thoroughly thriving activity, and it is evident that this will continue to be the case for a good many years to come.

July, 2005 *Roger Penrose*

Preface

Recent years have witnessed a tremendous improvement in the experimental verification of general relativity. Current experimental activities substantially outrange those of the past in terms of technology, manpower and, last but not least, money. They include earthbound satellite tests of weak-gravity effects, like gravitomagnetism in the Gravity-Probe-B experiment, as well as strong-gravity observations on galactic binary systems, including pulsars. Moreover, currently four large international collaborations set out to directly detect gravitational waves, and recent satellite observations of the microwave background put the science of cosmology onto a new level of precision.

All this is truly impressive. General relativity is no longer a field solely for pure theorists living in an ivory tower, as it used to be. Rather, it now ranges amongst the most accurately tested fundamental theories in all of physics. Although this success naturally fuels the motivation for a fuller understanding of the computational aspects of the theory, it also bears a certain danger to overhear those voices that try to point out certain, sometimes subtle, deficiencies in our mathematical and conceptual understanding. The point being expressed here is that, strictly speaking, a theory-based prediction should be regarded as no better than one's own structural understanding of the underlying theory. To us there seems to be no more sincere way to honor Einstein's "annus mirabilis" (1905) than to stress precisely this – his – point!

Accordingly, the purpose of the 319th WE-Heraeus Seminar "Mathematical Relativity: New Ideas and Developments", which took place at the Physikzentrum in Bad Honnef (Germany) from March 1 to 5, 2004, was to provide a platform to experts in Mathematical Relativity for the discussion of new ideas and current research, and also to give a concise account of its present state. Issues touching upon quantum gravity were deliberately not included, as this was the topic of the 271st WE-Heraeus Seminar in 2002 (published as Vol. 631 in the LNP series). We broadly categorized the topics according to their mathematical habitat: (i) differential geometry and differential topology, (ii) analytical methods and differential equations, and (iii) numerical methods. The seminar comprised invited one-hour talks and contributed half-hour talks. We are glad that most of the authors of the one-hour talks followed our invitation to present written versions for this volume.

We believe that the account given here is representative and of a size that is not too discouraging for students and non-experts.

Last but not least we sincerely thank the Wilhelm-and-Else-Hereaeus-Foundation for its generous support, without which the seminar on Mathematical Relativity would not have been possible and this volume would not have come into existence.

Tübingen - Freiburg - Berlin
July, 2005

Jörg Frauendiener
Domenico Giulini
Volker Perlick

Table of Contents

Part III Numerical Methods

List of Contributors

Maria Babiuc
Department of Physics
and Astronomy University
of Pittsburgh
Pittsburgh, PA 15260
USA
maria@einstein.phyast.pitt.edu

Robert Beig
Institut für Theoretische
Physik der Universität Wien
Boltzmanngasse
1090 Wien, Austria
robert.beig@univie.ac.at

Beverly K. Berger
Physics Division
National Science Foundation
Arlington, VA 22207
USA
bberger@nsf.gov

Sergio Dain
Max-Planck-Institut für
Gravitationsphysik
Am Mühlenberg 1, 14476 Golm
Germany
dain@aei.mpg.de

Paul E. Ehrlich
Department of Mathematics
University of Florida
Gainesville, FL 32611-8105
USA
ehrlich@math.ufl.edu

José L. Flores
Department of Mathematics
Stony Brook
University Stony Brook
NY 11794-3651
USA
and
Permanent address:
Departamento de Álgebra
Geometría y Topología
Universidad de Málaga
Campus Teatinos
29071 Málaga, Spain
floresj@math.sunysb.edu
floresj@agt.cie.uma.es

Simonetta Frittelli
Department of Physics
Duquesne University
Pittsburgh
PA 15282, USA
simo@mayu.physics.duq.edu

Roberto Gómez
Pittsburgh Supercomputing Center
4400 Fifth Avenue
Pittsburgh
PA 15213, USA
gomez@psc.edu

Luis Lehner
Department of Physics
& Astronomy
Brigham Young University
Provo, UT 84602
USA
lehner@lsu.edu

Robert J. Low
Mathematics Group
School of MIS, Coventry University
Priory Street
Coventry CV1 5FB
U.K.
mtx014@coventry.ac.uk

Antonio Masiello
Dipartimento di Matematica
Politecnico di Bari
Via Amendola 126/B
Bari 70125
Italy
masiello@poliba.it

Dave Neilsen
Department of Physics & Astronomy
Brigham Young University
Provo, UT 84602
USA
Theoretical Astrophysics 130-33
California Institute of Technology
Pasadena, CA 91125
USA
dneils1@lsu.edu

Alan D. Rendall
Max-Planck-Institut für
Gravitationsphysik
Am Mühlenberg 1
14476 Golm
Germany
rendall@aei.mpg.de

Miguel Sánchez
Departamento de Geometría y
Topología, Facultad de Ciencias
Universidad de Granada
Avenida Fuentenueva s/n
E-18071 Granada, Spain
sanchezm@ugr.es

Olivier Sarbach
Department of Physics
& Astronomy
Louisiana State University
Baton Rouge
LA 70803
USA
and
Theoretical Astrophysics 130-33
California Institute of Technology
Pasadena, CA 91125
USA
sarbach@phys.lsu.edu

László B. Szabados
Research Institute for Particle
and Nuclear Physics
Hungarian Academy of Sciences
1525 Budapest 114, P. O. Box 49
Hungary
lbszab@rmki.kfki.hu

Béla Szilágyi
Department of Physics
and Astronomy
University of Pittsburgh
Pittsburgh, PA 15260
USA
bela@einstein.phyast.pitt.edu

Manuel Tiglio
Department of Physics
& Astronomy
Louisiana State University
Baton Rouge, LA 70803
USA
and
Center for Computation
and Technology
Louisiana State University
Baton Rouge, LA 70803
USA

and
Center for Radiophysics
and Space Research
Cornell University, Ithaca
NY 14853
tiglio@phys.lsu.edu

Jeffrey Winicour
Department of Physics
and Astronomy
University of Pittsburgh
Pittsburgh, PA 15260
USA

and
Max-Planck-Institut für
Gravitationsphysik
Am Mühlenberg 1
14476 Golm, Germany
jeff@einstein.phyast.pitt.edu

Part I

Differential Geometry and Differential Topology

A Personal Perspective on Global Lorentzian Geometry

Paul E. Ehrlich

Department of Mathematics, University of Florida, Gainesville, FL 32611-8105, USA
ehrlich@math.ufl.edu

Dedicated to Professor John K. Beem on the occasion of his retirement from the University of Missouri–Columbia.

Abstract. A selected survey is given of aspects of global space-time geometry from a differential geometric perspective that were germane to the First and Second Editions of the monograph **Global Lorentzian Geometry** and beyond.

1 Introduction

Any student of Riemannian geometry is exposed to a wonderful global result and basic working tool, which goes back to Hopf and Rinow [55]. If (N, g_0) is a Riemannian manifold, then an associated Riemannian distance function $d_0 : N \times N \to \mathbb{R}$ is given by

$$d_0(p,q) = \inf\{L(c) \mid c : [0,1] \to N \text{ is a piecewise smooth curve with } c(0) = p,\ c(1) = q\}\,. \tag{1}$$

Then the promised result guarantees the equivalence of the following conditions:

Theorem 1. (Hopf–Rinow Theorem) *For any Riemannian manifold (N, g_0), the following are equivalent:*

(1) **metric completeness***: (N, d_0) is a complete metric space;*
(2) **geodesic completeness***: for any $v \in TN$, the geodesic $c_v(t)$ in N with initial condition $c_v{}'(0) = v$ is defined for all values of an affine parameter t;*
(3) *for some point $p \in N$, the exponential map exp_p is defined on all of T_pN;*
(4) **finite compactness***: every subset K of N that is d_0-bounded has compact closure.*

Moreover, if any one of (1) *through* (4) *holds, then (N, g_0) also satisfies*

(5) **minimal geodesic connectability***: given any $p, q \in N$, there exists a smooth geodesic segment $c : [0,1] \to N$ with $c(0) = p$, $c(1) = q$, and $L(c) = d_0(p,q)$.*

P.E. Ehrlich: *A Personal Perspective on Global Lorentzian Geometry*, Lect. Notes Phys. **692**, 3–34 (2006)
www.springerlink.com

Finally, the Heine–Borel property of basic topology implies (via (4)) that *all* Riemannian metrics for a compact manifold are automatically complete. Also, the basic examples that one introduces in a beginning Riemannian geometry course, such as S^n, $\mathbb{R}P^n$, $\mathbb{R}^n$, and T^n, all carry complete Riemannian metrics as usually described. From one viewpoint, one may regard the Hopf–Rinow Theorem as asserting that complete Riemannian metrics are the proper objects of study in the global differential geometry of Riemannian manifolds.

In a somewhat related vein, thanks to (1) of the Hopf–Rinow Theorem, it is well known that if the space of all Riemannian metrics $\mathrm{Riem}(N)$ for a fixed smooth manifold N is considered, then

$$\begin{gathered}\text{both geodesic completeness and} \\ \text{geodesic incompleteness are } C^0\text{-stable in } \mathrm{Riem}(N)\ .\end{gathered} \tag{2}$$

An elementary proof may be found in [14].

Now if we leave the Riemannian world and enter the realm of General Relativity, then unlike the basic complete or compact examples explained in elementary Riemannian geometry courses, we first find that several basic examples such as Schwarzschild space-time and the big bang cosmological models are nonspacelike geodesically *incomplete*. Moreover, in the 1970s attention was primarily focused on noncompact manifolds, because any compact space-time contains a closed timelike curve, thus violating the basic chronology condition of General Relativity.

The Hopf–Rinow Theorem fails to hold in general space-times. Indeed, we have explicitly recalled the statement of this result because much research has been set in the arena of what can be rescued for space-times and, later, semi–Riemannian manifolds. (For example, Beem [4] concerned what could be done with finite compactness, especially in the globally hyperbolic case.)

Also, nothing as simple as (2) holds for the space $\mathrm{Lor}(M)$ of all Lorentzian metrics for a given smooth manifold M without imposing further conditions on the background space-time (M, g) in question.

As one basic aspect of the failure of the Hopf–Rinow Theorem for space-times, it should be emphasized at the outset that compactness of the underlying manifold M by itself does *not* imply geodesic completeness of the space-time (M, g). A well-known result in basic Riemannian geometry is the proof that a homogeneous Riemannian metric on an arbitrary smooth manifold is automatically geodesically complete. This result fails to hold for indefinite metrics, but in Marsden [60] it was noted that a compact space-time with a homogeneous metric is geodesically complete (providing an early example of some less naïve aspects of the Hopf–Rinow Theorem being valid for space-times).

Much later, in Carrière [26], it was shown that a compact, flat space-time is geodesically complete; thus, adding the requirement that the Riemannian curvature tensor vanish rescues that aspect of the Hopf–Rinow Theorem. A second, more recent example which may be cited is the result that a compact

Lorentzian manifold which admits a timelike Killing field is automatically geodesically complete, obtained first with the assumption of constant curvature in Kamishima [56] and in full generality in Romero and Sánchez [73].

To help the reader recall the time frame in which the First Edition came into being, the following Table 1 shows the dates of publication for various selected texts in differential geometry and General Relativity.

Table 1. Publication Dates for Selected Standard References in Differential Geometry and General Relativity

R. Penrose, Techniques of Topology in General Relativity	1972
S. Hawking and G. Ellis, The Large Scale Structure of Space-time	1973
C. Misner, K. Thorne, and J. Wheeler, Gravitation	1973
R. Sachs and H. Wu, General Relativity for Mathematicians	1977
J. Beem and P. Ehrlich, Global Lorentzian Geometry	1981
B. O'Neill, Semi-Riemannian Geometry	1983

2 Some Aspects of Limit Constructions

In the context of the Eberlein–O'Neill compactification for complete, noncompact Riemannian manifolds of nonpositive sectional curvature (cf. [30]) and applications to the differential geometry of the geodesic flow, continuity and limit properties of sequences of geodesics (often expressed in the language of the exponential map) were essentially employed. The following aspects of the underpinnings of these topics especially captivated the author during his graduate studies in Riemannian geometry. First, the routine use of the compactness of the unit sphere bundle over any compact subset of the Riemannian manifold (N, g_0), and particularly the compactness of the set

$$\{v \in T_pN \,:\, g_0(v,v) = 1\} \tag{3}$$

was noted. A second key geometric property was the existence of a minimal geodesic ray

$$\gamma : [0, +\infty) \to (N, g_0) \tag{4}$$

based at each point p of N, i.e., a half geodesic γ as above with $\gamma(0) = p$ and

$$L(\gamma|_{[0,t]}) = d_0(\gamma(0), \gamma(t)) \qquad \text{for all } t \geq 0\,. \tag{5}$$

Here one could take a sequence $\{q_n\}$ in N with $\lim d_0(p, q_n) = +\infty$, and let σ_n be a unit speed minimal geodesic with $\sigma_n(0) = p$ and $\sigma_n(d_0(p, q_n)) = q_n$. Then taking any convergent subsequence of $\{\sigma_n'(0)\}$, say, converging to the unit vector w in T_pN, the unique geodesic σ in N with initial condition $\sigma'(0) = w$ provided the desired ray based at p.

With this last construction in mind, it was only a small step to generalize to the important asymptotic geodesic construction. Given the ray γ based at p and any other point q in N, let $t_n \to +\infty$, and this time let σ_n be a unit speed minimal geodesic segment from q to $\gamma(t_n)$. Then considering

$$\{\sigma_n'(0)\} \quad \text{in } T_qN \tag{6}$$

and letting w be any limit vector of this sequence by the compactness of (3), the geodesic $\sigma(t) = \exp_q(tw)$, i.e., the unique geodesic $\sigma : [0, +\infty) \to (N, g_0)$ with initial condition $\sigma'(0) = w$, was minimal as a limit of minimal segments and was thus a geodesic ray based at q, said to be *asymptotic* to the given ray γ. In these Riemannian studies, the uniqueness of the asymptotic geodesic σ to γ was considered and also under various curvature hypotheses, it was desired to estimate $d_0(\gamma(t), \sigma(t))$ as $t \to +\infty$.

When space-times (M, g) rather than Riemannian manifolds are considered, an immediate road block to employing the above machinery is the failure of the set of unit timelike tangent vectors based at a point p of M to be compact (even though this set is closed). A further difficulty is that a sequence

$$\{v_n \in T_pM \,:\, \|v_n\| = -1\} \tag{7}$$

of unit timelike tangent vectors cannot converge to a null vector n, even though the set of noncompact directions in T_pM is itself compact, so that

$$\text{direction}(v_n) \to \text{ direction}(n) \tag{8}$$

is indeed possible.

If any point p in a space-time (M, g) is selected, then emanating from p we have the three families of timelike, spacelike, and null geodesics. Especially in view of the limit arguments summarized above in global Riemannian geometry, since the null geodesics are, in a naïve, imprecise fashion, limits of the spacelike and timelike geodesics, it might be hoped (as was once hoped in General Relativity) that, possibly, continuity arguments could be obtained for the different types of geodesic completeness. For example, perhaps timelike and spacelike geodesic completeness (or incompleteness) might force null geodesic completeness (or incompleteness). However, in a series of examples, these earlier hopes were found to be too optimistic. Kundt [57] gave an example which was timelike and null geodesically complete, but not spacelike complete. Geroch [46] gave a globally hyperbolic example, conformal to Minkowski 2-space, which was null and spacelike complete, but not timelike complete.

Before discussing Beem's contribution [6] to this topic, we discuss a wonderful way in which semi-Riemannian manifolds differ from definite metric manifolds, from the viewpoint of the geodesic equation. When considering perturbations of Riemannian and semi-Riemannian metrics, attention is often restricted to *conformal changes* of the metric. If $\Omega : M \to (0, +\infty)$ is a given smooth function, then

$$\overline{g} = \Omega^2 \, g \tag{9}$$

is called a conformal change of metric. (The factor of 2 produces pleasant curvature and connection formulae.) Note that in the space-time case, the null, timelike and spacelike tangent vectors for both g and $\overline{g}$ are the same; hence the basic causality conditions like chronological, strongly causal, globally hyperbolic, etc., hold for both (M, g) and $(M, \overline{g})$ simultaneously.

A second aspect, absent for the definite case for which there are no null vectors, is that null geodesics for (M, g) remain null pregeodesics for $(M, \overline{g})$. To see this, write the conformal factor in the form

$$\overline{g} = e^{2f} \, g \, . \tag{10}$$

Then the Levi–Civita connections ∇ and $\overline{\nabla}$ for (M, g) and $(M, \overline{g})$ are related by

$$\overline{\nabla}_X Y = \nabla_X Y + X(f)Y + Y(f)X - g(X, Y)\mathrm{grad}(f) \, . \tag{11}$$

Especially, if X is a null vector field on M (so that $g(X, X) = 0$), then

$$\overline{\nabla}_X X = \nabla_X X + 2X(f)X \, . \tag{12}$$

Thus, if β is a null geodesic on (M, g) and $\nabla_{\beta'}\beta' = 0$, it follows that

$$\overline{\nabla}_{\beta'}\beta' = 2\beta'(t)(f)\beta'(t) \tag{13}$$

and it is known that if such an equation holds, then β may be reparametrized to be a null geodesic of $(M, \overline{g})$. For Riemannian manifolds, by contrast, the last gradient term in equation (11) will *not* vanish, for any nonzero $X = Y$, if the gradient is nonzero, and hence geodesics never persist under general conformal changes of metric.

With this background established, we can report that in [6] an example was given employing a conformal change involving an infinite product of factors. These factors were supported on suitably chosen subsets of a globally hyperbolic space-time to preserve timelike and spacelike geodesic completeness, but to produce null geodesic incompleteness for the deformed space-time metric. Hence, combining [6] with the examples of these earlier authors, no two types of geodesic completeness (or incompleteness) imply the third type. Beem liked to phrase this as follows:

Theorem 2. *Timelike geodesic completeness, null geodesic completeness, and spacelike geodesic completeness are logically inequivalent.*

In a somewhat related area, Beem also studied another important issue in space-time geodesic geometry at the time. A well-known result of Nomizu and Ozeki [68] asserts that an *arbitrary* Riemannian metric for a smooth manifold can be made geodesically complete by a conformal change of metric. On the other hand, the situation for space-times had been seen to be more complicated. Misner [61] gave a 2-dimensional null geodesically incomplete

example which could not be made complete by any conformal change (since the incomplete null geodesics were future trapped in a compact set and would remain future trapped null pregeodesics under any conformal change of metric). Seifert [77] showed that if (M, g) were stably causal, then a conformal change of metric could be made which would produce future nonspacelike geodesic completeness. Clarke [29] showed that a strongly causal space-time could be made null geodesically complete. So that is the setting for Beem's paper [5] on "Conformal changes and geodesic completeness."

In this paper, Beem formulated what he termed "Condition N" for non-imprisonment:

Definition 1. *The causal space-time* (M, g) *will satisfy* Condition *N if, for each compact subset* K *of* M*, there is no future inextendible nonspacelike curve* $x(t)$ *which is totally future imprisoned in* K*.*

Here it should be noted that while in Riemannian geometry attention is often simply restricted to *geodesics*, typically in General Relativity one has to consider all nonspacelike *curves*, not just geodesics. Here also the curve $x(t)$ is said to be *totally future imprisoned in* K if there exists t_1 so that $x(t) \in K$ for all $t \geq t_1$.

By making a sequence of conformal changes related to a compact exhaustion and taking the infinite product of those functions for the final conformal factor, Beem established.

Theorem 3. *Let* (M, g) *be a causal space-time which satisfies Condition N. Then there is some conformal factor* Ω *such that* $(M, \Omega^2 g)$ *is null and timelike geodesically complete.*

As a corollary, it followed that if (M, g) were distinguishing, strongly causal, stably causal or globally hyperbolic, then Condition N held, so that (M, g) could be made nonspacelike geodesically complete by a conformal change of metric. In Beem and Powell [22] an interesting study was made of Condition N for doubly warped products.

Since conformal changes are being considered, we will briefly summarize the first collaboration of Beem and the author in [8], "Conformal deformations, Ricci curvature and energy conditions on globally hyperbolic spacetimes." The author's thesis research was originally motivated by efforts to better understand a result of T. Aubin [1], which considered the question of deforming a Riemannian metric of nonnegative Ricci curvature and all Ricci curvatures positive at some point, to a metric of everywhere positive Ricci curvature.

Looking at [1] led to the study of *local convex deformations*: conformal deformations of a given Riemannian metric expressed in terms of the distance to the boundary of a convex metric ball. It was found that if the given Ricci curvature was nonnegative, then positive Ricci curvature could be produced in an annular region of the boundary of the convex metric ball, cf. [31]. Since

the distance from a point prior to the cut locus is nicely related to the index form, rather precise and detailed calculations and estimates could be made.

In Beem and Ehrlich [8], this situation was studied for globally hyperbolic space-times, where the situation was found to be rather more intricate. Since the intrinsic metric balls given by the Lorentzian distance function are noncompact and generally go off to infinity (cf. Fig. 4.4 in [15]), the analogous intrinsic construction of Ehrlich [31] could not be employed. Instead, in [8], a convex normal neighborhood B centered at p with local coordinates $x = (x_1, x_2, \ldots, x_n)$ was employed, and the auxiliary (nonintrinsic) distance function

$$f(p) = \sum_i x_i(p)^2 \tag{14}$$

was used to construct deformations with support in B. As might be expected, given this combination of intrinsic and auxiliary geometries, it was a more technical problem to calculate and study the Ricci curvature of the deformed metric. In contrast to the nicer Riemannian situation, where positive Ricci curvature was produced in a whole annular neighborhood of the boundary of the convex ball, it was found in the relativistic setting that positive Ricci curvature could only be guaranteed near the "north polar cap."

3 The Lorentzian Distance Function and Causal Disconnection

Even if it has been ten years or more since the reader has been the recipient of a graduate course in Riemannian geometry, she or he will no doubt still recall the pleasant properties that for a complete Riemannian manifold (N, g_0), the Riemannian distance function d_0 is continuous, and moreover, the metric topology induced by d_0 coincides with the given manifold topology. What is even more remarkable and perhaps less often remembered is that these properties are equally valid for an arbitrary *incomplete* Riemannian metric. Furthermore, it is taken for granted that $d_0(p, q)$ is finite for all $p, q \in N$.

Let (M, g) be an arbitrary space-time and let p, q be two points of M. If there is no future directed nonspacelike curve from p to q, set $d(p, q) = 0$; if there *is* such a curve, let

$$\begin{aligned} d(p,q) = \sup\{L(c) \,|\, c : [0,1] \to M \text{ is a piecewise smooth} \\ \text{future directed nonspacelike curve with } c(0) = p \text{ and } c(1) = q\} \,. \end{aligned} \tag{15}$$

Then this defines what some authors term the *Lorentzian distance function*

$$d = d(g) : M \times M \to [0, +\infty] \tag{16}$$

and other more physically motivated authors term *proper time*. (Note that unlike the Riemannian case, (15) does not bound the values of $d(p, q)$ from

above by $L(c)$ for any selected curve c.) When Beem and the author surveyed the scene after completing [8], it seemed that time was ripe for a more systematic exposition of the properties and uses of the space-time distance function. It had received some discussion in the monographs already published (see Table 1), and in some research papers. But confusion was found between nonspacelike conjugate points and nonspacelike cut points in certain aspects of the timelike index theory as explained in Hawking and Ellis [54] (indeed, nonspacelike cut points had not yet been formulated in the literature). An intrinsic Morse index theorem for null geodesic segments in arbitrary space-times had not yet been published, despite several works such as Uhlenbeck [78] and Woodhouse [80], among others.

Working some of these issues out was accomplished in Beem and Ehrlich [9, 10, 11], and in greater detail in the First Edition of **Global Lorentzian Geometry** [12]. In place of the complete metric of Riemannian geometry, what emerged was an interesting interplay between the causal properties of the given space-time and the continuity (and other properties) of the Lorentzian distance function. Philosophically, this aspect emerges since $d(p,q) > 0$ iff $q \in I^+(p)$. For example, at the one extreme of *totally vicious space-times*, the Lorentzian distance always takes on the value $+\infty$ (cf. p. 137 in [15]). Less drastically, if (M,g) contains a closed timelike curve passing through p, then $d(p,q) = +\infty$ for all $q \in I^+(p)$. In an allied vein, a space-time (M,g) is chronological iff its distance function vanishes identically on the diagonal

$$\Delta(M) = \{(p,p) \,:\, p \in M\} \,. \tag{17}$$

In general, the Lorentzian distance function is only lower semi-continuous. The strength of the additional property of upper semi-continuity is seen in the result that if a distinguishing space-time has a continuous distance function, then it is causally continuous.

At the other extreme from totally vicious space-times in the hierarchy of causality are globally hyperbolic space-times. In some sense, these space-times share many of the properties of complete Riemannian manifolds. For instance, the Lorentzian distance function of a globally hyperbolic space-time is both continuous and finite-valued. (Indeed, it may be shown that a strongly causal space-time (M,g) is globally hyperbolic iff all Lorentz metrics g' in the conformal class $C(M,g)$ also have finite-valued distance functions $d(g')$.) Second, for globally hyperbolic space-times, Seifert [76] and others had established the important working tool of *maximal nonspacelike geodesic connectability*: given any $p, q \in M$ with $p \leqslant q$, there exists a nonspacelike geodesic segment $c : [0,1] \to (M,g)$ with $c(0) = p$, $c(1) = q$, and $L(c) = d(p,q)$, in exact analogy with property (5) of the Hopf–Rinow Theorem stated in Sect. 1. In view of the last several properties of the distance function, the theories of the timelike and null cut loci were studied for globally hyperbolic space-times in Beem and Ehrlich [9].

Now let us turn to the main aim of this section, the discussion of causally disconnected space-times. Here is a general pattern which is common in global Riemannian geometry:

$$\begin{gathered}\text{complete Riemannian metric}\\ \text{and}\\ \text{curvature inequality}\\ \text{implies}\\ \text{topological or geometric conclusion.}\end{gathered} \tag{18}$$

A celebrated early example of (18) is the Topological Sphere Sphere Theorem of Riemannian geometry of the 1950s and 1960s: suppose that a complete, simply connected Riemannian manifold admits a metric whose sectional curvatures vary between 1/4 and 1, but are always strictly greater than 1/4. Then the manifold M must be homeomorphic to the standard round sphere of the same dimension as the given manifold.

In contrast with the Riemannian situation, we have already remarked in Sect. 1 that many standard examples of space-times *fail* to be nonspacelike geodesically complete. As we thought about (18) and certain of the singularity theorems already published in General Relativity from the viewpoint of differential geometry, the following contrasting pattern (19) emerged:

$$\begin{gathered}\text{curvature inequality (gravitation is attractive)}\\ \text{and}\\ \text{physical or geometric assumption}\\ \text{implies}\\ \text{the existence of an incomplete timelike or null geodesic.}\end{gathered} \tag{19}$$

As we were working on the concept of causal disconnection, we also had as a motivation that some cosmological and physical models are not globally hyperbolic, but are only strongly causal. Thus we wanted to establish a formalism in the more general strongly causal setting for which (a) the space-time distance function may be less tractable than for globally hyperbolic space-times, and also (b) the tool of maximal nonspacelike geodesic connectability is *not* available. (Later, in the context of the Lorentzian splitting theorem in the timelike geodesically complete case, Newman [63] and Galloway and Horta [44] would find the use of almost maximizers essential.) We were also motivated by the theory of the end structure for Riemannian or topological manifolds in establishing this concept. Strong causality did at least have the virtue that convergence in the limit curve sense and convergence in the C^0 topology on curves were closely related, as well as the fact that the upper semi-continuity of arc length in the C^0 topology fits in well with the lower semi-continuity of the space-time distance function (cf. Beem, Ehrlich and Easley [15], Sect. 3.3). Hence, the limit curve apparatus already established in General Relativity served as a partial substitute for some of the geodesic limit constructions for complete Riemannian manifolds reviewed at the beginning of Sect. 2. As the concept was stated in [15], p. 283:

Definition 2. *A space-time* (M, g) *is said to be* causally disconnected by a compact set K *if there exist two infinite sequences* $\{p_n\}$ *and* $\{q_n\}$ *diverging to infinity such that for each* n, $p_n \leqslant q_n$, $p_n \neq q_n$, *and all future directed nonspacelike curves from* p_n *to* q_n *meet* K.

This definition as finally formulated had the virtue that all Lorentzian metrics in the conformal class $C(M, g)$ are causally disconnected if any one of them is, unlike the original formulation in Beem and Ehrlich [7] which also assumed the finite distance condition

$$0 < d(p_n, q_n) < +\infty \quad \text{for all } n\,. \tag{20}$$

This finiteness assumption made it possible to use a simpler limit curve construction procedure, but unfortunately this earlier version of causal disconnection was only conformally invariant in the case that (M, g) was globally hyperbolic, so condition (20) was later dropped. As pointed out on Fig. 8.1 of [15], non-globally hyperbolic space-times may be causally disconnected.

By using local distance functions related to a compact exhaustion $\{B_n\}$ of M, almost maximizers related to the local distance functions, and taking limits (thanks to strong causality), the following result was obtained.

Theorem 4. *Let* (M, g) *denote a strongly causal space-time which is causally disconnected by a compact set* K. *Then* M *contains a nonspacelike geodesic line which intersects* K.

Here, a *line* is a past and future inextendible nonspacelike geodesic which realizes the Lorentzian distance function between every pair of its points. Now, unlike the familiar situation in the global Riemannian geometry of complete Riemannian manifolds, *no* assertion is being made here that the line is geodesically complete, only that it is inextendible.

With this result in hand, a singularity theorem fitting the pattern of (19) could be obtained:

Theorem 5. *Let* (M, g) *be a chronological space-time of dimension greater than or equal to three which is causally disconnected. If* (M, g) *satisfies the timelike convergence condition and the generic condition, then* (M, g) *is nonspacelike geodesically incomplete.*

Lurking in the background is a well-used result in causality theory of General Relativity that if (M, g) is a chronological space-time such that each inextendible null geodesic has a pair of conjugate points, then (M, g) is strongly causal (cf. [15], p. 467). In the statement of Theorem 5, we find two of the curvature conditions traditionally imposed in this branch of General Relativity. The first, the *timelike convergence condition*, is simply stated as $\mathrm{Ric}(v, v) \geq 0$ for all timelike tangent vectors v (hence, by continuity, the same condition holds for all nonspacelike tangent vectors). The second condition is somewhat more mysterious to differential geometers, and on a first pass can be paraphrased as the assertion that every inextendible nonspacelike geodesic

has some suitable nonzero sectional curvature. Indeed, formulating this condition in various ways easier for differential geometers to understand was done in Beem and Ehrlich [12] and Beem, Ehrlich and Easley [15]; (cf. Beem and Parker [21] among others for a discussion of the physical aspects of this condition).

In any event, the generic condition in the more unfriendly language of tensor calculus is discussed in the author's second favorite passage in Hawking and Ellis (54], p. 101), where K denotes the tangent vector to the null geodesic under consideration:

> "As in the timelike case, this condition will be satisfied for a null geodesic which passes through some matter provided that the matter is not pure radiation (energy-momentum tensor type II of §4.3) and moving in the direction of the geodesic tangent vector K. It will be satisfied in empty space if the null geodesic contains some point where the Weyl tensor is non-zero and where K does not lie in one of the directions (there are at most four such directions) at that point for which $K^c K^d K_{[a} C_{b]cd[e} K_{f]} = 0$. It therefore seems reasonable to assume that in a physically realistic solution every timelike or null geodesic will contain a point at which $K^a K^b K_{[c} R_{d]ab[e} K_{f]}$ is not zero. We shall say that a space-time satisfying this condition satisfies the *generic condition*."

So from the cynical viewpoint, one might take the following interpretation from this paragraph, which the author did for many years in perfect contentment:

$$\text{physically realistic} \implies \text{generic condition}\,. \tag{21}$$

On the other hand, "generic" has a precise meaning in differential geometry and topology; a condition is said to hold *generically* when it holds on an open, dense subset of the space in question. It never occurred to the author to ponder how (21) interfaced with this more precise definition of "generic," so he was thus delighted in the early 1990s to receive two preprints from J. Beem and S. Harris, published as [17, 18], the first with the especially charming title "The generic condition is generic." Since the most precise results in these papers are a bit complicated to state, we will content ourselves here with just giving four of the more easily stated results obtained in these two publications:

(1) $Ric(w, w) \neq 0$ implies w is generic.
(2) All vectors in T_pM are nongeneric implies that the curvature tensor at p vanishes identically.
(3) Constant curvature implies all null vectors are nongeneric.
(4) If (M^4, g) does not have constant sectional curvature at p, then the generic null directions at p form an open dense subset of the two-sphere of all null directions at p.

Thus, one could interpret (4) as stating that "the generic condition is generic after all," since a physically realistic universe should probably not have constant sectional curvature.

4 The Stability of Geodesic Completeness Revisited

In the First Edition of **Global Lorentzian Geometry** [12], a short Sect. 6.1 was written, entitled "Stable Properties of $\mathrm{Lor}(M)$ and $\mathrm{Con}(M)$," which was partly inspired by results of Lerner [58]. A motivation for this type of investigation in General Relativity had been provided by the hypotheses in the Singularity Theorems. If a condition held on an open subset of metrics in the space $\mathrm{Lor}(M)$ of all Lorentzian metrics for a given smooth manifold M, then philosophically a robuster theorem would result since this part of the hypotheses would remain true under suitable perturbations of the given metric, desirable since measurements cannot be made with infinite precision.

A result quoted in this Sect. 6.1 was the C^r-stability of geodesic completeness in $\mathrm{Lor}(M)$ for all $r \geq 2$. In the remaining two sections of Chap. 6, based on [13], a question raised in [58] was studied – the stability of timelike and null geodesic incompleteness for Robertson-Walker space-times. Here, a *Robertson–Walker space-time* was taken to be a warped product

$$M = (a, b) \times_f H \tag{22}$$

where (H, h) was a homogeneous Riemannian manifold and the metric tensor had the form

$$g = -dt^2 + fh \; . \tag{23}$$

That is how matters stood until 1985, when a copy of P. Williams' Ph.D. thesis [79], "Completeness and its stability on manifolds with connection," was received unexpectedly in the mail. This article revealed that there was a significant gap in the previous arguments for the C^r-stability of geodesic completeness in $\mathrm{Lor}(M)$, and that in fact neither geodesic completeness nor geodesic incompleteness was C^r-stable, although a stronger topology could be placed on $\mathrm{Lor}(M)$ which made geodesic completeness stable.

From a certain perspective, a good deal of research in global space-time geometry during the next decade can be viewed as trying to understand the more complicated geometry of the space of geodesics, once it was realized that Proposition 6.4 on page 175 of [12] failed to be valid.

In Williams [79], explicit studies were made of the system of null geodesics on the 2-torus T^2 which has background flat metric $g = dx\, dy$. For the first example, Williams studied the sequence of metrics

$$g_n = dx\, dy + \left(\frac{\sin x}{n}\right) dy^2 \tag{24}$$

and observed that $x = 0$ represents an incomplete null geodesic on (T^2, g_n) for all n. Hence, null geodesic completeness fails to be C^r-stable. For the second example, Williams considered

$$g_n = dx\,dy + (1 - \cos x + 1/n)dy^2 \tag{25}$$

and observed that while

$$dx\,dy + (1 - \cos x)dy^2 \tag{26}$$

contains an incomplete null geodesic, the metrics g_n are all null geodesically complete. Hence, null geodesic incompleteness fails to be C^r-stable.

In a series of papers, of which we will only discuss results from Beem and Parker [19, 20], the concept of *pseudoconvex geodesic system* was formulated, (cf. Parker [70] for the PDE motivation for this work). The theory was first formulated for metric connections and later broadened to linear connections. Here is how the concept is defined for the nonspacelike geodesics of a space-time (M, g).

Definition 3. *The space-time* (M, g) *is* causally pseudoconvex *iff for each compact subset* K *of* M*, there is a compact subset* K' *of* M *such that if* $\gamma : [a, b] \to M$ *is a nonspacelike geodesic with* $\gamma(a) \in K$ *and* $\gamma(b) \in K$*, then* $\gamma([a, b])$ *is contained in* K'*.*

As well as being a convexity statement akin to taking the convex hull of a set, this condition can be understood as a kind of internal completeness condition. It rules out incompleteness arising, for example, by taking a causal diamond $M = I^+(p) \cap I^-(q)$ and deleting a single point.

A second condition Beem and Parker imposed was that neither end of any of the geodesics in the geodesic system should be totally imprisoned in a compact set.

Definition 4. *The space-time* (M, g) *is* causally disprisoning *if, for each inextendible nonspacelike geodesic* $\gamma : (a, b) \to M$ *and any* $t_0 \in (a, b)$*, both sets* $\{\gamma(t) : a < t \leq t_0\}$ *and* $\{\gamma(t) : t_0 \leq t < b\}$ *fail to have compact closure.*

It is interesting that while neither causal pseudoconvexity nor causal disprisonment is a stable property by itself, *together* causal pseudoconvexity and causal disprisonment are C^1-stable in $\mathrm{Lor}(M)$. Indeed, the combination of these two properties may be regarded as a generalization of global hyperbolicity, for which Geroch [48] observed the C^0-stability in $\mathrm{Lor}(M)$.

In place of the assumption of Riemannian completeness, Beem and Parker [20] proved the following working tool for a manifold M with linear connection ∇.

Lemma 1. *Let* (M, ∇) *be both pseudoconvex and disprisoning. Assume that* $p_n \to p$ *and* $q_n \to q$ *for distinct* p*,* q *in* M*. If each pair* p_n*,* q_n *can be joined by a geodesic segment, then there exists a geodesic segment from* p *to* q*.*

Using this tool, Beem and Parker obtained a result akin to the type of thing classically obtained in Riemannian geometry for Cartan–Hadamard manifolds.

Theorem 6. *Let (M, ∇) be both pseudoconvex and disprisoning. If (M, ∇) has no conjugate points, then (M, ∇) is geodesically connected. Thus for each p in M the exponential map $\exp_p : T_pM \to M$ is a diffeomorphism of M with $\mathbb{R}^n$.*

Here is an example of how these two conditions under consideration rescue the stability of geodesic completeness.

Theorem 7. *Let (M, g) be causally pseudoconvex and causally disprisoning. If (M, g) is nonspacelike geodesically complete, then there is a fine C^1-neighborhood $U(g)$ of g in Lor(M) such that all g' in $U(g)$ are nonspacelike geodesically complete.*

In Beem and Ehrlich [14], a more conceptual study of certain of the constructions in Williams [79] was made. First, in a partial return to the roots of Ehrlich [31], a study was made of how conformal changes interfaced with null geodesic completeness. It was found that "small" conformal changes will destroy neither null completeness nor null incompleteness for pseudo-Riemannian manifolds (note that (24) and (25) are not conformal changes of metric). As a consequence, it follows that for a compact manifold M, all metrics in the conformal class $C(M, g)$ are either null geodesically complete or null geodesically incomplete. In the 1990s, Romero and Sánchez at Granada and others conducted studies of the geodesic behavior on compact spacetimes, obtaining a detailed and rich understanding (cf. [73] and [75] for two examples out of many).

Secondly, in the spirit of the Williams' examples which contained closed null geodesics, the following general result was obtained:

Theorem 8. *Let (M, g) be a pseudo-Riemannian manifold with a closed null geodesic $\beta : [0, 1] \to M$ satisfying $\beta'(0) = \beta'(1)$. Then each C^∞-fine neighborhood $U(g)$ of g in Pseudo(M) contains a metric g_1 which contains an incomplete closed null geodesic (and thus is null incomplete).*

The proof relies on a very non-Riemannian phenomenon important in certain aspects of General Relativity: given a smooth closed null geodesic $\gamma : [0, 1] \to (M, g)$, it is *not* automatically the case, as it is for Riemannian (timelike or spacelike) geodesics, that $\gamma'(0) = \gamma'(1)$, forcing the geodesic to be complete. Instead, all that can be guaranteed in the null case is that $\gamma'(0)$ and $\gamma'(1)$ are proportional. If these two vectors are unequal, then the given null geodesic is either future incomplete or past incomplete (cf. [15], pages 243–244 for a proof). So in [12], the given complete closed null geodesic β was perturbed in a tubular neighborhood (by a study of the Christoffel symbols and the geodesic ODEs) to a reparametrized null geodesic γ, with the same image as β, but having $\gamma'(1) = c\gamma'(0)$ with $c > 1$. Hence the new null geodesic γ was future incomplete.

5 The Lorentzian Splitting Problem

During the academic year 1979–1980, a Special Year in Differential Geometry was held at the Institute for Advanced Study in Princeton, New Jersey, with lead organizer Professor Shing–Tung Yau. We were fortunate enough to have been invited to participate in this program and elected to spend the second semester at the Institute. In the waning days of this session, Yau delivered a series of lectures, suggesting problems in differential geometry worthy of consideration. The list was published a few years later in Yau [81] in the Annals of the Mathematics Studies volume stemming from the Special Year in Differential Geometry at the Institute. As the author was attending these lectures, as a student of Professor Detlef Gromoll at Stony Brook and also having come under the influence of Professor Jeff Cheeger, he could not help but notice one of the problems Yau proposed:

Conjecture 1. (Yau) Show that a space-time (M, g) which is timelike geodesically complete, obeys the timelike convergence condition, and contains a complete timelike line, splits as an isometric product $(\mathbb{R} \times V, -dt^2 + h)$.

This problem was stated without motivation then as a proposal to obtain the space-time analogue of the celebrated Cheeger–Gromoll splitting theorem for Riemannian manifolds (cf.[28]). As the author thought about suggesting to Professor Beem that we attack this problem with the aid of a visiting postdoctoral researcher from Denmark, Dr. Steen Markvorsen, he was puzzled as to why Yau had formulated the problem with the hypothesis of timelike geodesic completeness rather than global hyperbolicity. For, recall from Sect. 3 that timelike geodesic completeness does *not* guarantee the existence of maximal timelike geodesic segments between chronologically related pairs of points, while global hyperbolicity *does* guarantee that helpful property. But this question was to go unanswered for several months until Professor G. Galloway, passing through Columbia for a short visit on the way back to Miami from a sabbatical in San Diego, could enlighten us himself as to what he had learned from S.-T. Yau.

At the time we began consideration of the problem, we were aware of various results on related issues employing maximal hypersurface methods, (cf. Bartnik [2], Gerhardt [45], and Galloway [41], among others). As decided nonexperts in maximal hypersurfaces, it seemed to us that it might be a wiser course to try to begin the study of the Busemann function of a timelike geodesic ray, since that tool had been a key ingredient rediscovered by Cheeger and Gromoll for use in their proof of the Riemannian splitting theorem. The author had studied Busemann functions during his stay in Bonn in connection with manifolds of negative curvature, and of course Beem was a student of Busemann himself, familiar with the text Busemann [25].

Since we had decided to take the course of action of exploring the Busemann function of a timelike geodesic ray, we returned once again to Cheeger

and Gromoll [28], which we had always found a challenging paper to understand. With the great emphasis on ellipticity of the Laplacian, while the d'Alembertian of General Relativity is hyperbolic, it also seemed like a daunting task to make any of this transform to the space-time case. Fortunately, just before we began our studies, we received an unexpected preprint in the mail which was published later as Eschenburg and Heintze [38]. At a first glance, it looked like an approach to the splitting problem that had a good chance of adapting to the space-time setting, and so with Markvorsen we set to work.

Many standard elementary methods in basic Riemannian geometry, in the context of constructing asymptotic geodesics as recalled in Sect. 2, rely on the compactness of the set of unit vectors based at a given point in the manifold. For space-times, we have recalled, however, the set of future unit timelike vectors based at a point (while closed) is noncompact and that a sequence of unit timelike tangent vectors can *never* converge to a null vector n.

By analogy with the Riemannian construction, let

$$\gamma : [0, +\infty) \to (M, g) \tag{27}$$

be a unit timelike geodesic ray, i.e., suppose that

$$L(\gamma|_{[0,t]}) = d(\gamma(0), \gamma(t)) \quad \text{for all } t \geq 0 \,. \tag{28}$$

Take any p in M with p in the chronological past of γ and any sequence $s_n > 0$ with $s_n \to +\infty$. Assuming that (M, g) is globally hyperbolic, construct unit speed maximal timelike geodesic segments c_n from p to $q_n = \gamma(s_n)$. As discussed above, in the Riemannian case, one could turn to the sequence of unit vectors $\{c_n'(0)\}$ and extract a convergent subsequence to define an asymptotic geodesic c to γ starting at p. But as we have indicated above, in the space-time case such a convergence is not guaranteed.

However, already at hand is the limit curve machinery for strongly causal space-times mentioned in Sect. 3. So instead, one may let c be a *nonspacelike limit curve* of the timelike geodesic segments $\{c_n\}$. Then two issues which must be dealt with are: (i) why is c timelike rather than null, and (ii) why is c future complete?

Inspired by the somewhat more general approach taken to the asymptotic geodesic construction in Busemann [25] (cf. Busemann [24]) for apparently the first appearance of what would later be termed by others the Busemann function), in Beem, Ehrlich, Markvorsen, and Galloway [16] the following definition was adopted for the concept of a nonspacelike asymptotic geodesic ray in which the point x corresponding to the point p above was allowed to vary in the limit construction:

Definition 5. *A* future co-ray *to γ from x will be a causal curve starting at x which is future inextendible and is the limit curve of a sequence of maximal*

length timelike geodesic segments from x_n *to* $\gamma(r_n)$ *for two sequences* $\{x_n\}$, $\{r_n\}$ *with* $x_n \to x$ *and* $r_n \to +\infty$.

To cope with the technicalities discussed above, the concept of the *timelike co-ray condition* was also formulated.

Definition 6. *The globally hyperbolic space-time* (M, g) *satisfies the* timelike co-ray condition *for the timelike line* $\gamma : (-\infty, +\infty) \to (M, g)$ *if, for each* x *in* $I(\gamma) = I^+(\gamma) \cap I^-(\gamma)$, *all future and past co-rays to* γ *from* x *are timelike.*

Here the analytic definition of the Busemann function corresponding to the future timelike geodesic ray $\gamma|_{[0,+\infty)}$ is given by:

$$(b_\gamma)^+(x) = \lim_{r\to\infty} (r - d(x, \gamma(r))) \, . \tag{29}$$

As mentioned in Sect. 3, the space-time distance function is generally less tractable than the Riemannian distance function. Hence, even issues such as continuity of (29) are less obvious. However, it was established in [16] that the timelike co-ray condition implied the continuity of the Busemann functions on $I(\gamma)$. Moreover, making the stronger hypothesis that all timelike sectional curvatures were nonpositive, it was established that the timelike co-ray condition holds on all of $I(\gamma)$, so that each of b^+, b^-, and $B = b^+ + b^-$ is continuous on $I(\gamma)$. Thanks to the aid of the powerful Toponogov Theorem for globally hyperbolic space-times with nonpositive timelike sectional curvatures, established in Harris [50, 51], it was also possible to prove that all past and future timelike co-rays to the given timelike geodesic line were complete. Hence, under the timelike sectional curvature hypothesis rather than the more desirable Ricci curvature hypothesis, one had what we liked to think of as "large scale control of the geometry on all of $I(\gamma)$." From this, one could obtain the splitting of $I(\gamma)$ as a metric product

$$(I(\gamma), g) = (\mathbb{R} \times H, -dt^2 + h) \tag{30}$$

where (H, h) was any level set of the Busemann function in the induced metric. (In Riemannian geometry, the corresponding level sets are called "horospheres.") Finally, by inextendibility arguments, one deduced that $I(\gamma) = M$.

What are some geometric issues hidden in the proofs involved in the $B = b^+ + b^-$ theory? Let γ be a complete timelike line as above and let $p \in I(\gamma)$. Form a future timelike co-ray c_1 to $\gamma|_{[0,+\infty)}$ and form a past timelike co-ray c_2 to $\gamma|_{[0,-\infty)}$, both starting at p. Then the biggest geometric issue is, why does it happen that

$$c_1{}'(0) = -c_2{}'(0) \tag{31}$$

so that c_1 and c_2 join together at p to form a smooth geodesic? Secondly, why is the geodesic globally maximal? Once these things have been established, then one can view the factor $\mathbb{R}$ of the splitting as being formed geometrically

by the collection of all of these asymptotic past and future rays to γ fitting together properly and H as any level set of the Busemann function.

We now briefly summarize how the proof of the splitting theorem was extended from the sectional curvature hypothesis to the desired timelike convergence condition that $\mathrm{Ric}(v,v) \geq 0$ for all timelike (hence all nonspacelike) tangent vectors. J.-H. Eschenburg [37] obtained the first important breakthrough in realizing that instead of trying for global control of the timelike co-rays on $I(\gamma)$ as in Beem, Ehrlich, Markvorsen, and Galloway [16], it was sufficient to obtain a splitting in a tubular neighborhood of the given timelike geodesic line and then extend the splitting to all of (M,g) through a procedure similar to the one used in making analytic continuation type arguments in complex analysis. Hence, in [37], the splitting theorem was obtained under the assumption of both timelike geodesic completeness and global hyperbolicity in the Ricci curvature case.

Working with this new idea, Galloway was able to remove the hypothesis of geodesic completeness shortly thereafter (cf. Galloway [42]). Then Newman returned to the original question of Yau and obtained the splitting for timelike geodesic completeness rather than global hyperbolicity (cf. Newman [63]). Here, Newman had to confront the issue that maximal nonspacelike geodesic segments could not be constructed without global hyperbolicity, so he had to work with almost maximizers instead of geodesics, introducing a higher level of complexity. A philosophy which emerged is that the existence of a maximal geodesic segment implies that things work out better in a tubular neighborhood of this maximal segment, in terms of the behavior of almost maximizers, the Busemann function, etc. In Galloway and Horta [44], these ideas were given a much simplified exposition and especially the original *timelike co-ray condition* of the earlier work [16] morphed into the *generalized timelike co-ray condition.*

Out of all of these results, the Lorentzian Splitting Theorem emerged.

Theorem 9. *Let (M,g) be a space-time of dimension $n \geq 3$ which satisfies each of the following conditions:*

(1) *(M,g) is either globally hyperbolic or timelike geodesically complete.*
(2) *(M,g) satisfies the timelike convergence condition.*
(3) *(M,g) contains a complete timelike line.*

Then (M,g) splits isometrically as a product $(\mathbb{R}\times V, -dt^2+h)$, where (H,h) is a complete Riemannian manifold.

Later, a version of the Splitting Theorem would also be obtained for maximal null geodesic lines, cf. Galloway [43].

As mentioned above, we had no inkling as to why Yau had proposed the problem of proving a Lorentzian splitting theorem (which was not mentioned by Yau in his lectures at the Institute or the Problem List written from those lectures), but that was explained to us by Galloway when he spoke in Columbia about what was published as [41]. Yau's motivation had been the

idea that timelike geodesic completeness should interface with the concept of "curvature rigidity" which had been formulated during the 1960s and 1970s in global Riemannian geometry, cf. especially the exposition in the introduction to the text Cheeger and Ebin [27], where it was first widely publicized.

Recall our earlier statement of the Sphere Theorem of global Riemannian geometry; if a complete, simply connected Riemannian manifold has sectional curvatures strictly 1/4-pinched, then it is homeomorphic to the n-sphere of the same dimension. In the statement of this result, there is a curvature condition of *strict* inequality. For curvature rigidity, the condition of strict inequality is relaxed to include the possibility of equality as well, and one tries to show that either the old possibility still obtains, or if it fails to be true, it fails in an *isometric* (hence "rigid") way.

Thus in the Riemannian example, if one relaxes the pinching on the sectional curvature to $\frac{1}{4} \leq K \leq 1$, then either the Riemannian manifold remains homeomorphic to the n-sphere (the old alternative), or if not, it is *isometric* to a symmetric space of rank one.

Already in Geroch [47], the idea had been presented that most space-times should be nonspacelike geodesically incomplete and also that a space-time should fail to be nonspacelike geodesically incomplete only under special circumstances (in the paragraph below, a white dot represents a geodesically complete space-time, a black dot a nonspacelike geodesically incomplete space-time):

> "Thus we expect that the diagrams for closed universes will be almost entirely black. There are, however, at least a few white points. There exist closed, geodesically complete, flat space-times... Perhaps there are a few other nonsingular closed universes, but these may be expected to appear either as isolated points or at least regions of lower dimensionality in an otherwise black diagram."

To see how the idea of curvature rigidity could apply in the context of timelike geodesic incompleteness, first let us state a simple prototype singularity theorem:

Theorem 10. *Let (M, g) be a space-time of dimension $n \geq 3$ which satisfies each of the following conditions:*

(1) *(M, g) contains a compact Cauchy surface.*
(2) *(M, g) satisfies $\mathrm{Ric}(v, v) \geq 0$ on all nonspacelike tangent vectors v.*
(3) *Every inextendible nonspacelike geodesic satisfies the generic condition.*

Then (M, g) contains an incomplete nonspacelike geodesic.

In this result, condition (2) *already* allows for equality, so that cannot be weakened. Thus, here curvature rigidity would call for dropping the requirement (3) of the generic condition that some curvature quantity is *nonzero* at some point of the geodesic. Hence, that is how the conjectured rigidity of

timelike geodesic completeness arises. This was apparently first published by one of Yau's Ph.D. students in Bartnik [3] as follows:

Conjecture 2. Let (M, g) be a space-time of dimension $n \geq 3$ which

(1) contains a compact Cauchy surface, and
(2) satisfies the timelike convergence condition $\mathrm{Ric}(v, v) \geq 0$ for all timelike v.

Then either (M, g) is timelike geodesically incomplete, or (M, g) splits *isometrically* as a product $(\mathbb{R} \times V, -dt^2 + h)$, where (H, h) is a compact Riemannian manifold.

The isometric splitting in the second alternative is precisely the manifestation of curvature rigidity here. The idea to solve this conjecture is a proof by contradiction. Suppose the space-time is *not* timelike geodesically incomplete. From the hypotheses, produce a nonspacelike line (recall Theorem 4 above), and prove that the line is timelike rather than null. Then under the assumption of timelike geodesic completeness the line is complete, so the Lorentzian Splitting Theorem may be applied to give the second alternative. Indeed, a result of this sort was obtained in [16] under the stronger sectional curvature hypothesis. A survey of later progress on this conjecture may be found in [15], Sect. 14.5.

6 Gravitational Plane Waves and the Nonspacelike Cut Locus

In August 1982 the author found himself in Gainesville, Florida, as a participant in an NSF CBMS Regional Conference with principal guest lecturer Professor Wilhelm Klingenberg of Bonn University. At that time, we were so impressed by the heat and humidity and the uniformity of the weather predictions on the evening newscast, we decided that we would never again set foot in the state of Florida. But in the fall of 1986, when we happened to be standing in the mathematics office at the University of Missouri, the Chair pointed out to us an advertisement which had just appeared in the Notices of the American Mathematical Society. This advertisement announced a building campaign for the Florida department with 20 new positions being added over the next five years, and during the first year of this process one of the desired fields called for a senior appointment in differential geometry, with the successful candidate also advising on the recruitment of several junior positions. We telephoned the new outside Chair at the University of Florida, Gerard Emch, and were later somewhat surprised to find ourselves back in Gainesville during December, 1986, on a job interview of several days. During this time in Gainesville, Emch showed us a copy of Penrose's paper [71], including an intriguing Fig. 2 on page 218 (which both Sánchez and the author

showed separately during our plenary lectures at this conference). Emch told us that he had done some preliminary calculations on the gravitational plane wave space-times inspired by this paper, but would like to work together on obtaining a more complete understanding if I were to indeed become the new Professor of differential geometry at the University of Florida.

Finally, during the spring semester, 1989, during the year that Professor Greg Galloway of the University of Miami was also in Gainesville, we three started meeting weekly in Emch's office for him to explain what he had learned from his prior studies of Penrose [71]. Soon, Emch and Galloway were eager to pass from this established arena and do some more exotic things like add dust to the basic model. Galloway even explained how Frankel's approach [40] to the Raychaudhuri equation was well adapted to such calculations. At the end of one of our sessions, it was decided that each of us should think of an aspect of this class of space-times on which our particular expertise could be brought to bear. Now the abstract theory of the nonspacelike cut locus had been developed for globally hyperbolic space-times in [9], and some standard elementary examples had been presented. However, nothing as exotic as the geodesic behavior in [71] had been considered, and also this class of space-times presented an initial challenge in that they all failed to be globally hyperbolic, so neither the tool of maximal nonspacelike geodesic connectability nor the general theory of the nonspacelike cut locus was immediately available. Thus we announced that we thought it would be interesting to understand the nonspacelike cut locus for this class of space-times, little suspecting that it would encompass four years to completely tie up all the loose ends, especially aspects of the achronal boundary (cf. [32, 33, 34, 35, 36]).

Since Sánchez in his lecture at the conference provided a complete discussion of plane fronted waves and other issues, we refer to his article in this volume (with Flores) for more general background and concentrate on the gravitational plane wave space-times and the nonspacelike cut and first conjugate loci in this section, cf. also the discussion in Chap. 13 of [15].

Let $\gamma : [0, a) \to (M, g)$ be an inextendible future directed nonspacelike geodesic in a general space-time (M, g) with Lorentzian distance function d given as in (15). Consider the condition

$$d(\gamma(0), \gamma(t_0)) = L(\gamma|_{[0,t_0)}) \tag{32}$$

for some t_0 in $[0, a)$ which particularly implies that $d(\gamma(0), \gamma(t_0))$ is finite. By the reverse triangle inequality for Lorentzian distance, $d(\gamma(0), \gamma(t))$ is finite for all $t \leq t_0$. Suppose that $\gamma|_{[0,t]}$ is not maximal for some t with $0 < t < t_0$. Then by definition of space-time distance, there exists a future causal curve $\sigma : [0, 1] \to (M, g)$ with $\sigma(0) = \gamma(0)$, $\sigma(1) = \gamma(t)$, and $L(\sigma) > L(\gamma|_{[0,t]})$. Forming the composition $\mu = \sigma \circ \gamma|_{[t,t_0]}$, we have $L(\mu) > L(\gamma|_{[0,t_0]}) = d(\gamma(0), \gamma(t_0))$, contradicting the definition of the distance (15). (For $t = 0$, if $d(\gamma(0), \gamma(0)) > 0$, then there is a closed timelike curve $\sigma : [0, 1] \to (M, g)$ with $\sigma(0) = \sigma(1) = \gamma(0)$ and traversing σ over and

over, we obtain $d(\gamma(0),\gamma(0)) = +\infty$, in contradiction.) Hence, if (32) holds for t_0, then

$$d(\gamma(0),\gamma(t)) = L(\gamma|_{[0,t]}) \tag{33}$$

holds for all t with $t \leq t_0$, e.g., the nonspacelike geodesic segment $\gamma|_{[0,t]}$ is *maximal* for all $t \leq t_0$.

Now in less tractable space-times, like the totally vicious space-times recalled in Sect. 3 with $d(p,q) = +\infty$ for all p, q in M, condition (32) never holds for any t_0. More benignly, if the chronological condition fails to hold at $\gamma(0)$, condition (32) fails to hold for any t_0. At the other extreme on the causality ladder, for globally hyperbolic space-times, the distance function is always finite valued and continuous. Thus condition (32) may be considered along any nonspacelike geodesic. Also, as recalled in prior sections, causally related pairs of points are joined by maximal geodesic segments. Hence, the concept of the nonspacelike cut locus fits most handily into the class of globally hyperbolic space-times, but may be formulated in more general space-times, cf. [9].

Returning to the inextendible future directed nonspacelike geodesic $\gamma : [0,a) \to (M,g)$ in the arbitrary space-time (M,g) and supposing that (33) holds for some $t \geq 0$, now set

$$t_0 = \sup\{t \in [0,a) \,:\, d(\gamma(0),\gamma(t)) = L(\gamma|_{[0,t]})\}\,. \tag{34}$$

Definition 7. *If $0 < t_0 < a$, then the point $\gamma(t_0)$ in (M,g) is said to be the* future nonspacelike cut point *of $p = \gamma(0)$ along γ.*

Consistent with our terminology above, in [12] and in several earlier journal articles, a future directed nonspacelike curve $\sigma : [0,d] \to (M,g)$ was said to be *maximal* if $L(\sigma) = d(\sigma(0),\sigma(d))$ and it was also noted that a maximal nonspacelike curve may be reparametrized as a smooth geodesic (cf. [15], p. 147). Employing this language, it may be checked (just as was done for the cut locus in the earlier Riemannian theory) that

(a) for $0 < s < t < t_0$, $\gamma|_{[s,t]}$ is the unique maximal nonspacelike geodesic segment in all of (M,g) between $\gamma(s)$ and $\gamma(t)$;
(b) $\gamma|_{[0,t]}$ is maximal for all t with $0 \leq t \leq t_0$;
(c) for all t with $t_0 < t < a$, there is a longer nonspacelike curve in (M,g) than $\gamma|_{[0,t]}$ between $\gamma(0)$ and $\gamma(t)$.

Especially, $\gamma|_{[0,t]}$ is maximal in the above sense up to and including the cut point $\gamma(t_0)$, but fails to be maximal past the cut point. Recalling the concept of timelike geodesic ray central in Sect. 5, if $\gamma : [0,a) \to (M,g)$ happens to be a timelike geodesic ray, then $t_0 = a$ in (34) and there is no cut point to $p = \gamma(0)$ along γ.

A more familiar concept is that of a nonspacelike conjugate point $\gamma(t_1)$ to $\gamma(0)$ along the nonspacelike geodesic γ, this roughly speaking being defined by

the existence of a nonzero smooth Jacobi vector field along γ which vanishes at both $t = 0$ and $t = t_1$. (The situation is a bit more technical in the case of a null geodesic, cf. [15], pp. 368–374.) Here, calculus of variation arguments are employed, especially in the timelike case, to show that for any t with $t_1 < t < a$, there is a 1-parameter family of future timelike curves from $\gamma(0)$ to $\gamma(t)$, each of which is longer than $\gamma|_{[0,t]}$ (cf. [15], p. 333). Additionally, all of the curves in this 1-parameter family may be taken to be "close" to the given geodesic segment $\gamma|_{[0,t]}$. (Thus, for example, in the Riemannian text [49], p. 121, the terminology "Nachbarkurve" is used which could be translated as "neighboring curve.") It is often written (cf. [54], p. 97) that at the conjugate point $\gamma(t_1)$, *infinitesimally* neighboring geodesics emanating from $\gamma(0)$ refocus or intersect at $\gamma(t_1)$. However, the geodesics need only refocus at $\gamma(t_1)$ *up to second order*, and thus there are not necessarily any geodesics emanating from $\gamma(0)$ which actually pass through $\gamma(t_1)$. Since the calculus of variations arguments show that past a nonspacelike conjugate point, longer neighboring curves join $\gamma(0)$ to $\gamma(t)$, it follows that the future cut point to $p = \gamma(0)$ along γ comes no later than the first future conjugate point to p along γ in either the timelike or the null geodesic cases.

An attractive aspect of the differential geometry of Jacobi fields and geodesics is a correspondence between Jacobi fields along the given timelike geodesic $\gamma : [0, a) \to (M, g)$ and variations of γ whose neighboring curves consist of timelike geodesics. In the context of the above paragraph, suppose that J is a smooth nonzero Jacobi field along γ with $J(0) = J(t_1) = 0$. Then if the traditional variation

$$\alpha(t, s) = \exp_p(t(\gamma'(0) + sJ'(0))) \tag{35}$$

is constructed, (cf. [15], Prop. 10.16), all the neighboring curves $t \to \alpha(t, s)$ are timelike geodesics issuing from $p = \gamma(0)$, and the variation vector field $V = \alpha_* \frac{\partial}{\partial s}|_{s=0}$ satisfies $V = J$, the given Jacobi field. Hence, $V(t_1) = J(t_1) = 0$. But since the smooth curve $s \to \alpha(t_1, s)$ is not a geodesic generally, the condition $V(t_1) = 0$ does *not* force $\alpha(t_1, s) = \gamma(t_1)$ for all s, e.g., the geodesic neighboring curves do not necessarily actually pass through $q = \gamma(t_1)$, whence the assertion neighboring geodesics "infinitesimally" refocus at q.

In the General Relativity literature (cf. [54], pp. 110–111]) a terminology has been used that a timelike geodesic segment $\gamma : [0, a) \to (M, g)$ is "maximal" if there is no point $\gamma(t)$, $t \in (0, a)$, which is conjugate to p along γ. To compare this concept geometrically with our usage of maximal, we are considering whether there is a *single* future directed nonspacelike curve σ from $\gamma(0)$ to $\gamma(t)$ which is longer than $\gamma|_{[0,t]}$, but with σ possibly very "far" from $\gamma|_{[0,t]}$. The definition given in [54] is considering whether there is a 1-parameter family σ_ϵ of future directed nonspacelike curves from $\gamma(0)$ to $\gamma(t)$, each of which lies in some tubular neighborhood of $\gamma|_{[0,t]}$ and each of which is longer than $\gamma|_{[0,t]}$.

Often the assumption of global hyperbolicity for a space-time acts as a good substitute for a complete metric in Riemannian geometry, even though

this paradigm is not always exact. In a somewhat startling result which contradicted erroneous arguments in all the standard textbooks, Margerin [59] gave examples to show that even for a compact Riemannian manifold, the first conjugate locus (i.e., the set of all first conjugate points along all geodesics issuing from a given point) need not be closed, even though elementary arguments do show that the cut locus of any point (i.e., the set of all cut points along all geodesics issuing from the given point) on a complete Riemannian manifold is always closed. Of course the timelike conjugate locus of a point in a space-time will generally not be closed, but because of the nonimprisonment property that a nonspacelike geodesic in a globally hyperbolic space-time must escape from any compact subset in finite affine parameter, it may be shown that the future (or past) first nonspacelike conjugate locus of any point in a globally hyperbolic space-time (M, g) is a closed subset of M, cf. [15], p. 315. Also, the geometric characterization of a cut point in a complete Riemannian manifold is faithfully mirrored for globally hyperbolic space-times. Let (M, g) be globally hyperbolic and let $q = \gamma(t_0)$ be the future cut point of $p = \gamma(0)$ along the timelike [respectively, null] geodesic segment γ from p to q. Then either one or possibly both of the following conditions hold:

(i) q is the first future conjugate point to p along γ, or
(ii) there exist at least two maximal timelike [respectively, null] geodesic segments from p to q.

Denote by $C_t^+(p)$ the future timelike cut locus of p in (M, g), i.e., the set of all timelike cut points along all future timelike geodesics issuing from p. Correspondingly, let $C_n^+(p)$ denote the future null cut locus of p in (M, g) consisting of all future null cut points along all null geodesics issuing from p, and define the future nonspacelike cut locus of p by

$$C^+(p) = C_t^+(p) \cup C_n^+(p) \; . \tag{36}$$

Employing alternatives (i) and (ii) above much as in the Riemannian theory, it may be established for globally hyperbolic space-times that the null cut locus and the nonspacelike cut locus of any point are closed subsets of M. It was recalled in Sect. 2, equation (13), that null geodesics remain null pregeodesics under conformal changes of the background space-time metric. (Since conformal deformations fail to preserve timelike geodesics, the behavior of timelike conjugate points or timelike cut points along a given timelike geodesic under conformal metric deformations cannot be considered.) Even though null conjugate points along a null geodesic will not remain invariant under conformal change of space-time metric, it is remarkable that elementary arguments involving the space-time distance function show that global conformal diffeomorphisms do preserve null cut points and hence the null cut locus, cf. [15], p. 308. This may be seen as a plus for the definition of "maximal" formulated using the space-time distance function and tying in

with the theory of the cut locus, instead of using the alternative definition recalled above employing the semi-definiteness of the index form and conjugate points.

Now as mentioned at the beginning of this section, neither the more general plane fronted waves nor gravitational plane waves themselves are globally hyperbolic. Hence, when we first saw Fig. 2 in Penrose [71] (reproduced in the article by Flores and Sánchez for these proceedings), while the question immediately came to mind as to whether the future null cut locus formed prior to the refocusing of the null geodesics at astigmatic conjugacy, it was not immediately clear how to resolve this issue since the abstract structure theory for the nonspacelike cut locus discussed above was only valid in the globally hyperbolic case. But the rich structure of the Killing fields and isometry groups of these space-times entailed that explicit calculations of the geodesics issuing from a suitably chosen point $P_0 = (0, 0, 0, u_0)$ and a mimicry of the Wronskian techniques of linear ODE theory gave sufficient insights into the general global geodesic behavior of gravitational plane wave space-times to compensate for the inapplicability of the abstract cut locus theory, cf. [33, 34, 35, 36].

Let $M = \mathbb{R}^4$ with global coordinates (y, z, v, u). Heuristically, one may think of starting with the usual coordinates (x, y, z, t) of Minkowski space-time and making the change of variables

$$u = \frac{1}{\sqrt{2}}(t - x), \qquad v = -\frac{1}{\sqrt{2}}(t + x) \tag{37}$$

where the second minus sign has been chosen so that $\partial/\partial v$ is past directed null. In these coordinates, the Minkowski metric has the form

$$\eta = 2du\, dv + dy^2 + dz^2 \,. \tag{38}$$

Definition 8. *A* gravitational plane wave *is the smooth manifold $M = \mathbb{R}^4$ equipped with a Lorentzian metric $\overline{g} = \eta + H(y, z, u)du^2$, where the function $H(y, z, u)$ has the quadratic form*

$$H(y, z, u) = f(u)(y^2 - z^2) + 2g(u)yz \tag{39}$$

where either $f(u)$ or $g(u)$ does not vanish identically.

In these geometries, the null hyperplanes

$$P(s) = \{(y, z, v, u) \in \mathbb{R}^4 \,:\, u = s\} \tag{40}$$

play a distinguished role, for the given space-time metric restricted to $P(s)$ is degenerate, and for any (y_0, z_0) in $\mathbb{R}^2$, the plane $P(s)$ contains the maximal null geodesic line (which happens also to be a straight line in the usual sense)

$$\beta(t) = (y_0, z_0, t, s) \tag{41}$$

which is therefore free of null conjugate or null cut points. But all other geodesics passing through $(y_0, z_0, 0, s)$ lying in $P(s)$ are straight lines which are also spacelike geodesics. The global coordinate $u : M \to \mathbb{R}$ plays a helpful role in understanding the global geometry and was termed a "quasi-time function" in [33, 34, 35, 36], cf. the article by Flores and Sánchez in these proceedings for a fuller discussion of the implications of this concept.

As stated above, because of the nature of the isometry group, general results may be deduced from explicit calculations based at $P_0 = (0, 0, 0, u_0)$. Equally well, results stated most simply without introducing a lot of notational apparatus in the polarized case $g(u) = 0$ are generally valid, so to simplify our discussion below we will also take $g(u) = 0$. A first wonderful consequence of the quadratic form of the metric (39), which fails for more general plane fronted waves, is that all members of this class of metrics are geodesically complete independent of the choice of $f(u)$ or $g(u)$. Hence to be technically precise, the behavior exhibited in Fig. 2 of Penrose [71] is not "singular" behavior, but rather "caustic" behavior. Also as a result of the quadratic form of the metric in (39), all share the property that the Ricci tensor vanishes, $\mathrm{Ric} = 0$, even though the metric will have nonzero curvature tensor (or equivalently, some nonzero sectional curvatures) unless $f(u) = g(u) = 0$ for all u. Thus recalling the discussion of the generic condition in Sect. 3, the singularity theorems of General Relativity imply the existence of what were termed "astigmatic conjugate pairs" $\{u_0, u_1\}$ in [33, 34, 35, 36]. Equivalently, this astigmatic conjugacy may be seen as an application of conjugacy theory in linear ODEs.

For the cover illustration of the Second Edition of Global Lorentzian Geometry [15], the two dimensional universal anti-de Sitter space-time was selected. This choice was made to emphasize that for space-times, unlike Riemannian manifolds, geodesic completeness (even in the presence of constant curvature) does not imply geodesic connectability. In the figure, points p and q were shown such that even though q is in the chronological future of p, there is an open neighborhood U of q in $I^+(p)$ such that no point x in U lies on any geodesic issuing from p, let alone a future timelike geodesic. Thus even in the presence of geodesic completeness, the exponential map of a space-time may fail to map onto open subsets of the space-time, whereas in the Riemannian case, geodesic completeness implies that given any q in (N, g_0), there exists a minimal geodesic segment from p to q, i.e., $\exp_p$ is onto.

Since geodesic connectability of the gravitational plane waves is under consideration, put

$$\begin{aligned}\mathrm{Conn}(P_0, u) = \{Q \in P(u) :\ & \text{there exists a geodesic} \\ & \text{from } P_0 = (0, 0, 0, u_0) \text{ to } Q\}\end{aligned} \tag{42}$$

with $P(u)$ the null hyperplane defined in (40). Then for a first astigmatic conjugate pair $\{u_0, u_1\}$ with $u_0 < u_1$ in $\mathbb{R}$, the following behavior was obtained by direct calculation:

(i) For Q with $u_0 < u(Q) < u_1$, there exists a unique geodesic from P_0 to Q and hence $\mathrm{Conn}(P_0, u) = P(u)$, whence $\dim(\mathrm{Conn}(P_0, u)) = 3$. Moreover, $P_0 \ll Q$ still with $u_0 < u(Q) < u_1$ iff P_0 is joined to Q by a maximal timelike geodesic segment. (Similarly, reflecting a basic result in General Relativity, $Q \in J^+(P_0) - I^+(P_0)$ iff P_0 is joined to Q by a maximal null geodesic segment.) Hence, there are no future nonspacelike cut points to P_0 prior to astigmatic conjugacy at $P(u_1)$.

(ii) At $u = u_1$, $\dim(\mathrm{Conn}(P_0, u_1)) = 2$ and as $\mathrm{Conn}(P_0, u_1)$ drops dimension by 1, every R in $\mathrm{Conn}(P_0, u_1)$ is conjugate to P_0 by a 1-parameter family of geodesics. (Recall our above remarks that the geodesics issuing from P_0 only need to refocus up to second order at conjugacy in general, but in this model the geodesics all actually refocus.)

A more concrete discussion of the timelike and null cut and conjugate loci is most easily given in the polarized case $g(u) = 0$ in (39), even though the structural results summarized here are valid generally. Again, explicit computations show that in this special case

$$\mathrm{Conn}(P_0, u_1) = P(u_1) \cap \{z = 0\} \, . \tag{43}$$

The first (future) null conjugate locus of P_0 in $\mathrm{Conn}(P_0, u_1)$ is a planar parabola in the plane (43) which separates the first spacelike conjugate locus from the first (future) timelike conjugate locus. (Particularly, every future null geodesic issuing from P_0 contains a conjugate point except for the maximal null geodesic of the form (41) with $y_0 = z_0 = 0$ lying in $P(u_0)$.) Note that while $\exp_{P_0}$ does not map onto $P(u_1)$, given any Q in $\mathrm{Conn}(P_0, u_1) = P(u_1) \cap \{z = 0\}$ and open neighborhood U of Q in M, there are points R in U which are joined to P_0 by geodesics; hence unlike the spacetime on the cover of [15], open sets of points are not omitted from the image of $\exp_{P_0}$ at astigmatic conjugacy.

As mentioned in paragraph (i) above, the nonspacelike cut locus of P_0 does not occur prior to astigmatic conjugacy at $P(u_1)$, settling the question we had wondered about during the spring of 1989. Also since a nonspacelike geodesic fails to be maximal past a conjugate point, the nonspacelike geodesics issuing from P_0 (apart from (41)) fail to be maximal past $u = u_1$ and hence the future timelike (respectively, null) cut locus of P_0 equals the first future timelike (respectively, null) conjugate locus of P_0 in $\mathrm{Conn}(P_0, u_1)$, cf. Fig. 13.1 on p. 489 of [15] for a sketch of the above discussion. Hence, the future null cut locus turned out to be a parabola in the plane $\mathrm{Conn}(P_0, u_1)$ and the future timelike cut locus consists of all points inside the parabola in this plane.

It has been tempting while working on[33, 34, 35, 36] to consider writing about how the gravitational plane waves just escape providing examples for many phenomena in General Relativity apart from their failure to be globally hyperbolic. From our viewpoint here, a good bit of this is explained by the failure of the single exceptional maximal null geodesic (41) in $P(u_0)$ to reach $P(u_1)$ unlike all the other geodesics issuing from P_0. In the article by Flores

and Sánchez in these proceedings, geodesic connectability for a wider class of space-times is studied by variational methods. From their viewpoint, the quadratic case (39) where a critical exponent is exactly equal to 2 is where the variational calculus breaks down, cf. also Flores and Sánchez [39] where it is shown that subquadratic growth does not preclude global hyperbolicity.

7 Some More Current Issues

From the viewpoint of the material discussed in this contribution, we find that among the most interesting areas for current research is the application of ideas from the Gromov–Hausdorff convergence theory of Riemannian geometry to the space-time setting. Here, J. Noldus spoke at this conference on "Lorentzian Gromov Hausdorff Theory," reporting on recent progress, cf. Bombelli and Noldus [23], and Noldus [64, 65, 66, 67]. It should be emphasized that as is often usual, more than a mere translation from the Riemannian to the Lorentzian setting is required. At the beginning of Sect. 2, we mentioned some concepts associated with the construction of the Busemann boundary for a complete Riemannian manifold of nonpositive curvature, as initiated by P. Eberlein and B. O'Neill [30]. A second interesting issue is the deployment of this concept in the space-time setting; here we refer to articles of S. Harris [52, 53] for a recent approach to this topic. Thirdly, the splitting theorem for null geodesic lines should be more widely understood, cf. Galloway [43] for the origin of this result. Finally, we mention as a more minor loose end related to Sect. 3, a conjecture of J. Beem that all space-times should be causally disconnected.

References

1. T. Aubin: Métriques riemannienes et courbure. J. Diff. Geom. **4**, 383–424 (1970) 8
2. R. Bartnik: Existence of maximal surfaces in asymptotically flat space-times. Commun. Math. Phys. **94**, 155–175 (1984) 17
3. R. Bartnik: Remarks on cosmological space-times and constant mean curvature surfaces. Commun. Math. Phys. **117**, 615–624 (1988) 22
4. J.K. Beem: Globally hyperbolic space-times which are timelike Cauchy complete. Gen. Rel. Grav. **7**, 339–344 (1976) 4
5. J.K. Beem: Conformal changes and geodesic completeness. Commun. Math. Physics **49**, 179–186 (1976) 8
6. J.K. Beem: Some examples of incomplete space-times. Gen. Rel. Grav. **7**, 501–509 (1976) 6, 7
7. J.K. Beem, P.E. Ehrlich: Distance lorentzienne finie et géodésiques f-causales incomplètes. C. R. Acad. Sci. Paris Ser. A **581**, 1129–1131 (1977) 12
8. J.K. Beem, P.E. Ehrlich: Conformal deformations, Ricci curvature and energy conditions on globally hyperbolic space-times. Math. Proc. Camb. Phil. Soc. **84**, 159–175 (1978) 8, 9, 10

9. J.K. Beem, P.E. Ehrlich: The space-time cut locus. Gen. Rel. Grav. **11**, 89–103 (1979) 10, 23, 24
10. J.K. Beem, P.E. Ehrlich: Cut points, conjugate points and Lorentzian comparison theorems. Math. Proc. Camb. Phil. Soc. **86**, 365–384 (1979) 10
11. J.K. Beem, P.E. Ehrlich: A Morse index theorem for null geodesics. Duke Math. J. **46**, 561–569 (1979) 10
12. J.K. Beem, P.E. Ehrlich: *Global Lorentzian Geometry* (Marcel Dekker, New York 1981) 10, 13, 14, 16, 24
13. J.K. Beem, P.E. Ehrlich: Stability of geodesic incompleteness for Robertson–Walker space-times. Gen. Rel. Grav. **13**, 239–255 (1981) 14
14. J.K. Beem, P.E. Ehrlich: Geodesic completeness and stability. Math. Proc. Camb. Phil. Soc. **102**, 319–328 (1987) 4, 16
15. J.K. Beem, P.E. Ehrlich, K.L. Easley: *Global Lorentzian Geometry*, 2nd edn (Marcel Dekker, New York 1996) 9, 10, 11, 12, 13, 16, 22, 23, 24, 25, 26, 28, 29
16. J.K. Beem, P.E. Ehrlich, S. Markvorsen, G. Galloway: Decomposition theorems for Lorentzian manifolds with nonpositive curvature. J. Diff. Geom. **22**, 29–42 (1985) 18, 19, 20, 22
17. J.K. Beem, S.G. Harris: The generic condition is generic. Gen. Rel. Grav. **25**, 939–962 (1993) 13
18. J.K. Beem, S.G. Harris: Nongeneric null vectors. Gen. Rel. Grav. **25**, 963–973 (1993) 13
19. J.K. Beem, P.E. Parker: Whitney stability of solvability. Pacific J. Math. **116**, 11–23 (1985) 15
20. J.K. Beem, P.E. Parker: Pseudoconvexity and geodesic connectedness. Annali Math. Pura Appl. **155**, 137–142 (1989) 15
21. J.K. Beem, P.E. Parker: Sectional curvature and tidal accelerations. J. Math. Phys. **31**, 819–827 (1990) 13
22. J.K. Beem, T. Powell: Geodesic completeness and maximality in Lorentzian warped products. Tensor N.S. **39**, 31–36 (1982) 8
23. L. Bombelli, J. Noldus: The moduli space of isometry classes of globally hyperbolic spacetimes. Class. Quantum Grav. **21**, 4429–4454 (2004). `gr-qc/0402049` 30
24. H. Busemann: Über die Geometrien, in denen die "Kreise mit unendlichem Radius" die kürzesten Linien sind. Math. Annalen **106**, 140–160 (1932) 18
25. H. Busemann: *The Geometry of Geodesics* (Academic Press, New York 1955) 17, 18
26. Y. Carrière: Autour de la conjecture de L. Markus sur les variétés affines. Invent. Math. **95**, 615–628 (1989) 4
27. J. Cheeger, D. Ebin: *Comparison Theorems in Riemannian Geometry* (North–Holland, Amsterdam 1975) 21
28. J. Cheeger, D. Gromoll: The splitting theorem for manifolds of nonnegative Ricci curvature. J. Diff. Geom. **6**, 119–128 (1971) 17, 18
29. C.J.S. Clarke: On the geodesic completeness of causal space-times. Proc. Camb. Phil. Soc. **69**, 319–324 (1971) 8
30. P. Eberlein, B. O'Neill: Visibility manifolds. Pacific J. Math. **46**, 45–109 (1973) 5, 30
31. P.E. Ehrlich: Metric deformation of curvature. I: Local convex deformations. Geom. Dedicata **5**, 1–23 (1976) 8, 9, 16
32. P. Ehrlich: Astigmatic conjugacy and achronal boundaries. In: *Geometry and Global Analysis*, ed by T. Kotake, S. Nishikawa, and R. Schoen (Tohoku University, Sendai, Japan 1993) pp 197–208 23
33. P. Ehrlich, G. Emch: Gravitational waves and causality. Reviews in Mathematical Physics **4**, 163–221 (1992). Errata **4**, 501 (1992) 23, 27, 28, 29

34. P. Ehrlich, G. Emch: Quasi-time functions in Lorentzian geometry. Lecture Notes in Pure and Applied Mathematics **144**, 203–212 (1992) 23, 27, 28, 29
35. P. Ehrlich, G. Emch: The conjugacy index and simple astigmatic focusing. Contemporary Mathematics **127**, 27–39 (1992) 23, 27, 28, 29
36. P. Ehrlich, G. Emch: Geodesic and causal behavior of gravitational plane waves: astigmatic conjugacy. Proc. Symp. in Pure Mathematics (Amer. Math. Soc.) **54**, Part 2, 203–209 (1993) 23, 27, 28, 29
37. J.-H. Eschenburg: The splitting theorem for space-times with strong energy condition. J. Diff. Geom. **27**, 477–491 (1988) 20
38. J.-H. Eschenburg, E. Heintze: An elementary proof of the Cheeger–Gromoll splitting theorem, Ann. Global Analysis Geometry **2**, 141–151 (1984) 18
39. J.L. Flores, M. Sánchez: Causality and conjugate points in general plane waves. Class. Quantum Grav. **20**, 2275–2291 (2003) 30
40. T. Frankel: *Gravitational Curvature* (W.H. Freeman, San Francisco 1979) 23
41. G. Galloway: Splitting theorems for spatially closed space-times. Commun. Math. Phys. **96**, 423–429 (1984) 17, 20
42. G. Galloway: The Lorentzian splitting theorem without completeness assumption. J. Diff. Geom. **29**, 373–387 (1989) 20
43. G. Galloway: Maximum principles for null hypersurfaces and splitting theorems. Ann. Henri Poincaré **1**, 543–567 (2000) 20, 30
44. G. Galloway, A. Horta: Regularity of Lorentzian Busemann functions. Trans. Amer. Math. Soc. **348**, 2063–2084 (1996) 11, 20
45. C. Gerhardt: Maximal H-surfaces in Lorentzian manifolds. Commun. Math. Phys. **96**, 523–553 (1983) 17
46. R.P. Geroch: What is a singularity in general relativity? Ann. Phys. (NY) **48**, 526–540 (1968) 6
47. R.P. Geroch: Singularities in relativity. In: *Relativity*, ed by M. Carmeli, S. Fickler, L. Witten (Plenum, New York 1970) pp 259–291 21
48. R.P. Geroch: Domain of dependence. J. Math. Phys. **11**, 437–449 (1970) 15
49. D. Gromoll, W. Klingenberg, W. Meyer: *Riemannsche Geometrie im Großen*, Lecture Notes in Mathematics **55** (Springer, Berlin 1968) 25
50. S.G. Harris: Some comparison theorems in the geometry of Lorentz manifolds. Ph.D. Thesis, University of Chicago (1979) 19
51. S.G. Harris: A triangle comparison theorem for Lorentz manifolds. Indiana Math. J. **31**, 289–308 (1982) 19
52. S. Harris: Topology of the future chronological boundary: universality for spacelike boundaries. Class. Quantum Grav. **17**, 551–603 (2000) 30
53. S. Harris: Boundaries on spacetimes: an outline. In: *Advances in differential geometry and general relativity. The Beemfest*, Contemporary Mathematics **359**, ed by S. Dostoglou, P.E. Ehrlich (American Mathematical Society, Providence 2004) pp 65–85 30
54. S.W. Hawking, G.F.R. Ellis: *The Large Scale Structure of Space-Time* (Cambridge University Press, Cambridge 1973) 10, 13, 25
55. H. Hopf, W. Rinow: Über den Begriff der vollständigen differentialgeometrischen Fläche. Comment. Math. Helv. **3**, 209–225 (1931) 3
56. Y. Kamishima: Completeness of Lorentz manifolds of constant curvature admitting Killing vector fields. J. Diff. Geom. **37**, 569–601 (1993) 5
57. W. Kundt: Note on the completeness of space-times. Zeitschrift für Physik **172**, 488–489 (1963) 6

58. D.E. Lerner: The space of Lorentz metrics. Commun. Math. Phys. **32**, 19–38 (1973) 14
59. C. Margerin: General conjugate loci are not closed. Proc. Symp. in Pure Mathematics (Amer. Math. Soc.) **54**, Part 3, 465–478 (1993) 26
60. J.E. Marsden: On completeness of homogeneous pseudo-Riemannian manifolds. Indiana Univ. Math. J. **22**, 1065–1066 (1973) 4
61. C.W. Misner: Taub–NUT space as a counterexample to almost anything. In: *Relativity and Astrophysics I: Relativity and Cosmology*, ed by J. Ehlers (American Mathematical Society, Providence 1967) pp 160–169 7
62. C.W. Misner, K. Thorne, J. A. Wheeler: *Gravitation* (W.H. Freeman, San Francisco 1973)
63. R.P.A.C. Newman: A proof of the splitting conjecture of S.-T. Yau. J. Diff. Geom. **31**, 163–184 (1990) 11, 20
64. J. Noldus: A new topology on the space of Lorentzian metrics on a fixed manifold. Class. Quantum Grav. **19**, 6075–6107 (2002) 30
65. J. Noldus: A Lorentzian Gromov Hausdorff notion of distance. Class. Quantum Grav. **21**, 839–850 (2004). `gr-qc/0308074` 30
66. J. Noldus: The limit space of a Cauchy sequence of globally hyperbolic spacetimes. Class. Quantum Grav. **21**, 851–874 (2004). `gr-qc/0308075` 30
67. J. Noldus: Lorentzian Gromov Hausdorff theory as a tool for quantum gravity kinematics. PhD Thesis, University of Genf (2004). `gr-qc/0401126` 30
68. K. Nomizu, H. Ozeki: The existence of complete Riemannian metrics. Proc. Amer. Math. Soc. **12**, 889–891 (1961) 7
69. B. O'Neill: *Semi-Riemannian Geometry with Applications to Relativity* (Academic Press, New York 1983)
70. P. Parker: Geometry of Bicharacteristics. In: *Advances in differential geometry and general relativity. The Beemfest*, Contemporary Mathematics **359**, ed by S. Dostoglou, P.E. Ehrlich (American Mathematical Society, Providence 2004) pp 31–40 15
71. R. Penrose: A remarkable property of plane waves in general relativity. Rev. Mod. Phys. **37**, 215–220 (1965) 22, 23, 27, 28
72. R. Penrose: *Techniques of Differential Topology in Relativity*, Regional Conference Series in Applied Math. **7** (Society for Industrial and Applied Mathematics, Philadelphia 1972)
73. A. Romero, M. Sánchez: On completeness of certain families of semi-Riemannian manifolds. Geom. Dedicata **53**, 103–117 (1994) 5, 16
74. R.K. Sachs, H. Wu: *General Relativity for Mathematicians* (Springer, New York 1977)
75. M. Sánchez: Structure of Lorentzian tori with a Killing vector field. Trans. Amer. Math. Soc. **349**, 1063–1080 (1997) 16
76. H.-J. Seifert: Global connectivity by timelike geodesics. Zeitschrift für Naturforschung **22a**, 1356–1360 (1967) 10
77. H.-J. Seifert: The causal boundary of space-times. Gen. Rel. Grav. **1**, 247–259 (1971) 8
78. K. Uhlenbeck: A Morse theory for geodesics on a Lorentz manifold. Topology **14**, 69–90 (1975) 10
79. P.M. Williams: Completeness and its stability on manifolds with connection. Ph.D. Thesis, University of Lancaster (1984) 14, 16

80. N.M.J. Woodhouse: An application of Morse theory to space-time geometry. Commun. Math. Phys. **46**, 135–152 (1976) 10
81. S.T. Yau: Problem Section in *Seminar on differential geometry*. Ann. of Math. Studies **102**, ed by S.T. Yau (Princeton University Press, Princeton 1982) pp 669–706

The Space of Null Geodesics (and a New Causal Boundary)

Robert J. Low

Mathematics Group, School of MIS, Coventry University, Priory Street, Coventry CV1 5FB, U.K.
mtx014@coventry.ac.uk

Abstract. The space of null geodesics, G, of a space-time, $\mathcal{M}$, carries information on various aspects of the causal structure $\mathcal{M}$. In this contribution, we will review the space of null geodesics, G, and some natural structures which it carries, and see how aspects of the causal structure of $\mathcal{M}$ are encoded there. If $\mathcal{M}$ is strongly causal, then G has a natural contact manifold structure, points are represented in G by smooth Legendrian S^2s, and the relationships between these S^2s reflect causal relationships between the points of $\mathcal{M}$. One can also attempt to pass in the opposite direction with the intention of constructing a space-time from a family of S^2s in G; this process suggests a means of attaching end-points to null geodesics of $\mathcal{M}$, and thereby constructing a causal boundary. We close by summarizing some open questions in this general area.

1 Introduction

In Newtonian physics, the structure of space-time is fairly straight-forward. Although there is no notion of absolute rest, there is a notion of absolute time, and it always makes sense to say whether two events are simultaneous, and if not, which of them occurs first. Simultaneity is an equivalence relation, and can be used to slice space-time up into surfaces in a canonical way. Whether the situation at one event, p, can influence that at another, q, is determined by which of them happens at the later time (unless they are simultaneous, in which case they are independent). However, although this is conceptually straight-forward, it suffers from a major drawback: namely, it does not agree with the available data.

In general relativity, on the other hand, we model space-time as a smooth differentiable manifold, $\mathcal{M}$, equipped with a smooth Lorentz metric, g. Quantities associated with the metric, for example the connection, the Riemann tensor, the Ricci tensor and others, have physical interpretations and can be used to develop physics in this setting, and make testable - and tested! - predictions, which to date have proven consistent with the available experimental data [1]. There is also considerable activity in developing and carrying out new experimental tests of effects which have so far been too subtle to detect; two programmes of particular importance are Gravity Probe B [2] and LIGO [3].

R.J. Low: *The Space of Null Geodesics (and a New Causal Boundary)*, Lect. Notes Phys. **692**, 35–50 (2006)
www.springerlink.com

However, the question of when information at one event can affect that at another is rather less straight-forward than in the case of Newton's universe. The Lorentz metric implicitly provides the answer to this question: for it determines whether a curve in space-time can describe the path of a material particle, or influence. But in the general case, the situation can be much more complicated than in the Newtonian one.

For example, one can find space-times which locally look perfectly acceptable, but have the property that a particle can have a closed space-time trajectory, i.e. it can travel in such a way that it meets its own past self. The first, and perhaps the most famous, solution of this type was discovered by Gödel [4]. Such space-times are philosophically problematic, and the subject of much debate [5]; they raise awkward questions about the nature of free will, or nonlocal constraints on initial data. But even if we simply exclude such awkward behaviour by fiat, and restrict our attention to space-times with more acceptable behaviour, the situation remains complicated. Indeed, if a metric is simply expressed in terms of coordinate patches, it may be a difficult task to check whether the space-time is causally acceptable.

But considerations of causal structure are further reaching than just providing a reason for rejecting certain space-times as unacceptable. One would also like to know whether apparently reasonable initial data has a reasonable time evolution, or does some kind of singular behaviour occur eventually? If so, is this singularity decently concealed or can it be naked [6]? What are the properties of the edge, or boundary, of space-time? Even stating such questions precisely requires a good deal of causal machinery.

So, let us recall the basic question: given two points, p and q in $\mathcal{M}$ are p and q causally related in the sense that a physical influence can propagate from one to the other? This causal structure is distinctly more primitive than the metric structure on $\mathcal{M}$, since space-times with different metrics may have the same causal structure. It is, on the other hand, inextricably bound up with the metric structure since two space-times have the same causal structure if and only if they have the same null geodesics. Indeed, so crucial to the causal structure are the null geodesics that one can take the null geodesics themselves as primitive objects, regard points of space-time as derived objects, and profitably study aspects of the causal structure of space-time in this context instead.

In the remaining sections of this contribution I will describe the space of null geodesics, and in particular its topological and geometric structure. We will see how notions of causality in space-time are reflected in this new setting, providing elegant reinterpretations of familiar ideas, and also a powerful way of considering the development of wave fronts. Finally, I will suggest a means of attaching endpoints to endless null geodesics to provide a new type of conformal boundary.

First, however, I will give a very brief review of some causal theory, establishing standard terminology and definitions. This material is developed in

depth in Penrose's lecture notes on differential topology [7] and (with slightly different conventions) in the classical text of Hawking and Ellis [8].

We denote space-time by $\mathcal{M}$ (and, by a standard abuse of notation, will use $\mathcal{M}$ when $\langle \mathcal{M}, g \rangle$ would be correct). The tangent bundle of $\mathcal{M}$ is $T\mathcal{M}$, with fibre $T_p\mathcal{M}$ at p, and the cotangent bundle is $T^*\mathcal{M}$ with fibre $T_p^*\mathcal{M}$ at p. The isomorphism beteen $T\mathcal{M}$ and $T^*\mathcal{M}$ provided by the metric g will be used freely. We will use the convention that the metric g has signature $(+, -, \ldots, -)$, and say that a vector $v \in T_p\mathcal{M}$ is timelike if $g_p(v, v) > 0$, causal if $g_p(v, v) \geq 0$, null if $g_p(v, v) = 0$ and spacelike if $g_p(v, v) < 0$. Unless otherwise stated, $\mathcal{M}$ will be four-dimensional.

Also, $\mathcal{M}$ is said to be time-orientable if there exists a continuous timelike vector field t on $\mathcal{M}$; we will always assume that $\mathcal{M}$ is time-orientable. This does not in fact require any significant loss of generality: any space-time which is not time-orientable has a time orientable double cover [9]. Clearly, if t is such a timelike vector field, so is $-t$. We arbitrarily choose one of these as determining the future direction. As a consequence, we can distinguish between future pointing causal vectors (whose inner product with t is positive) and past pointing ones (whose inner product with t is negative).

A smooth curve is timelike (future pointing) if its tangent vector is everywhere timelike (future pointing), and similarly for causal, null, future or past pointing, or spacelike.

If $p, q \in \mathcal{M}$, then q is in the chronological future of p, written $q \in I^+(p)$, if there is a timelike future pointing curve $\gamma : [0, 1] \to \mathcal{M}$ with $\gamma(0) = p$, and $\gamma(1) = q$; similarly, q is in the causal future of p, written $q \in J^+(q)$, if there is a future pointing causal curve from p to q. For any point, p, $I^+(p)$ is open; but $J^+(p)$ need not, in general, be closed. $J^+(p)$ is, however, always a subset of the closure of $I^+(p)$.

One can use properties of these two ordering relations to define a causal space in the absence of any notion of metric, and study causal structure in this more general setting [10]; more recently, similar ideas have been used in Sorkin's causal set program of quantum gravity [11].

In addition, $E^+(p) = J^+(p) \setminus I^+(p)$ is the future horismos of p, and is ruled by segments of null geodesics emanating from p. A null geodesic originating at p can leave $E^+(p)$ and enter $I^+(p)$ if it intersects another, or passes a conjugate point [8] (intuitively, a point where infinitesimally separated null geodesics starting at p cross one another). Then we have the inclusion $E^+(p) \subseteq \partial I^+(p) = \partial J^+(p)$. Denoting by $N^+(p)$ all those points lying on a future pointing null geodesic starting at p, we also have $E^+(p) \subseteq N^+(p)$. In general, neither of $N^+(p)$ or $\partial I^+(p)$ is a subset of the other.

If $K \subset \mathcal{M}$, then $D^+(K)$ is the set of all points p such that any past-endless causal curve through p intersects K. $D^+(K)$ is called the future domain of dependence of K, and is the region where physics is entirely determined by data on K (in the absence of material which allows faster-than-light effects

to propagate), since no material influence can reach any element of $D^+(K)$ without passing through K.

One can use the causal relations I^+ and J^+ to impose conditions on space-time. Indeed, Carter has shown [14] that there is an infinite hierarchy of distinct conditions, each implied by all its successors, which can be imposed in terms of causal relations. I will restrain myself to listing a few of immediate relevance. Each of the following conditions is implied by its successor.

1. If there is no point p such that $p \in I^+(p)$, $\mathcal{M}$ is said to satisfy the chronological condition.
2. A space-time $\mathcal{M}$ which has no point p with a non-degenerate causal curve which starts and ends at p is said to satisfy the causal condition.
3. If each point p has arbitrarily small neighbourhoods which any causal curve intersects in a single component, $\mathcal{M}$ satisfies the condition of strong causality.
4. If $\mathcal{M}$ is causal and remains causal under small changes of g, it is stably causal.
5. If $\mathcal{M}$ is strongly causal and $J^\pm(p)$ is the topological closure of $I^\pm(p)$ for every $p \in \mathcal{M}$ (so $E^\pm(p) = \partial I^\pm(p)$, then $\mathcal{M}$ is causally simple.
6. If there is a spacelike surface $\mathcal{S}$ (i.e. a surface of codimension one whose tangent plane at each point contains only spacelike vectors) which every endless causal curve intersects in exactly one point, $\mathcal{M}$ is globally hyperbolic, and is the topological product of $\mathcal{S}$ with $\mathbb{R}$. $\mathcal{S}$ is called a Cauchy surface for $\mathcal{M}$.

We note that all of these concepts depend only on the conformal class of g, i.e. g may be replaced by Ωg, where Ω is a strictly positive function on $\mathcal{M}$, without any effect on causal properties.

2 Space of Null Geodesics

Even from the brief review above, it is clear that null geodesics are fundamental to the causal structure of $\mathcal{M}$. Motivated by this observation, we can consider the space of all null geodesics, and in particular investigate the relationships between its topology and geometry and the causal structure of $\mathcal{M}$.

In the following development, we will use the cotangent bundle of $\mathcal{M}$; for some purposes, it would be more natural to use the tangent bundle and the geodesic flow on the tangent bundle. Indeed, this approach has been used in the study of the space of geodesics of a Riemannian manifold [12] and in the more general case of the space of geodesics of a manifold with affine connection [13]. However, in the section after this one I wish to make use of some structures which naturally arise on the cotangent bundle, and so will work with the cotangent bundle from the beginning. As mentioned above, free

use will be made of the isomorphism which g gives between the tangent and cotangent bundles. We will now consider how the cotangent bundle structure provides natural structure on the space of null geodesics.

So let $T^*\mathcal{M}$ be the cotangent bundle of $\mathcal{M}$, and $\pi : T^*\mathcal{M} \to \mathcal{M}$ the canonical projection. There are two vector fields on $T^*\mathcal{M}$ of interest.

Let $\alpha \in T^*\mathcal{M}$, and define $f : \mathbb{R} \to T^*\mathcal{M}$ by $f(t) = t\alpha$. Then Δ, the Euler field, is defined by $\Delta(\alpha) = f_*(\partial/\partial t)(1)$, or, more concretely, if α has coordinates (q^i, n_i) then $\Delta(\alpha)$ is $n_i \partial/\partial n_i$.

The other vector field we require is the geodesic vector field, X_G. To define this on $T^*\mathcal{M}$ we first define the Hamiltonian function $H : T^*\mathcal{M} \to \mathbb{R}$ which sends each covector to its squared length given by g, and then X_G is the corresponding Hamiltonian vector field determined by $i_{X_G}(\omega) = -dH$. In terms of the usual coordinates,

$$X_G = n^i \frac{\partial}{\partial q^i} - \Gamma^i_{jk} n^j n_i \frac{\partial}{\partial n_k}$$

where indices are raised and lowered using g, and Γ^i_{jk} is the usual Christoffel symbol.

If $c : \mathbb{R} \to \mathcal{M}$ is a smooth curve, given in coordinates by $c(t) = q^i(t)$, then it has a natural lift to $T^*\mathcal{M}$ given by $(q^i(t), n_i(t))$, where $n_i = g_{ij}\dot{q}^i$. This is an integral curve of X_G iff c is an affinely parameterised geodesic in $\mathcal{M}$.

Now, we can restrict our attention to $N^*\mathcal{M}$, the subset of $T^*\mathcal{M}$ given by the future pointing null vectors (excluding the zero vector at each point). Since each of Δ and X_G are tangent to $N^*\mathcal{M}$, we can regard them as vector fields on this manifold. Furthermore, since the two are never linearly independent, the vector space spanned by X_G and Δ at each point gives a two-dimensional distribution on $N^*\mathcal{M}$, i.e. a two-dimensional subspace of $T(N^*\mathcal{M})$ at each point. In addition, since $[\Delta, X_G] = X_G$, this distribution is integrable, i.e. $N^*\mathcal{M}$ is foliated by two-surfaces whose tangent surfaces are the subspaces spanned by X_G and Δ at each point [15]. We can therefore consider the quotient space of integral surfaces.

It is perhaps more geometrically intuitive to construct this quotient space in stages. First, we can take the quotient space of integral curves of Δ in $N^*\mathcal{M}$, resulting in the bundle of future pointing null directions, $\mathbb{P}N^*\mathcal{M}$. Now, although X_G itself does not descend to this quotient space, the one-dimensional distribution of subspaces spanned by X_G does, and we again obtain a distribution. This time, the integral curves are the lifts of null geodesics in $\mathcal{M}$ to the bundle of null directions of $\mathcal{M}$, and the quotient space is obtained by identifying points on the same (lifted) null geodesics. We therefore call this quotient space the space of null geodesics, and denote it by $\mathcal{N}$.

Alternatively, we can take the quotient of $N^*\mathcal{M}$ under the action of X_G, to obtain the space of scaled null geodesics, and then take the further quotient which corresponds to forgetting the scaling.

Now that $\mathcal{N}$ has been provided with a topology, we can consider convergence of a sequence of null geodesics as curves in space-time. So let γ_n be a sequence of points in $\mathcal{N}$, and denote by Γ_n the corresponding curves in $\mathcal{M}$. Suppose $\gamma \in \mathcal{N}$. Then, since a neighbourhood of γ in $\mathcal{N}$ is the image under the projection from $\mathbb{P}N^*\mathcal{N}$ of a neighbourhood of a point on the lift of Γ to $\mathbb{P}N^*\mathcal{N}$, we see that $\gamma_n \to \gamma$ if there is a sequence of points $p_n \in \Gamma_n$ and a point $p \in \Gamma$ such that $p_n \to p$, and the tangent direction to Γ_n at p_n tends to the tangent direction to Γ at p. More naively, two null geodesics are close if they pass close to each other and the tangent directions are also close.

Note that this is not the topology we obtain by insisting that, for any neighbourhood U of Γ, each Γ_n eventually lies inside U - it may well be that for each Γ_n there are points which are very far from Γ.

Example 1. Let $\mathcal{M}$ be the Minkowski space, with the usual coordinates (t, x, y, z), and let γ_n be the null geodesic through the origin with tangent $(1, \cos(1/n), \sin(1/n), 0)$. Then the limit null geodesic has tangent $(1, 1, 0, 0)$, but the distance between the points where Γ_n and Γ intersect the surface $t = T$ can be made arbitrarily large by taking T large enough.

Furthermore, although Frobenius' theorem guarantees the existence of integral surfaces, and hence of a quotient space which inherits a topological structure, this need not in general be a manifold.

Example 2. Let $\mathbb{M}^2$ be two-dimensional Minkowski space, with the usual coordinates (t, x), and let $\mathcal{M}$ be obtained from $\mathbb{M}^2$ by identifying t with $t + 1$ and x with $x + \sqrt{2}$ for all x, t. Then the space of null geodesics, $\mathcal{N}$ has two components: $\mathcal{L}$, the space of left directed null geodesics, and $\mathcal{R}$, the space of right directed ones. All right directed null geodesics are parallel, and each is dense in $\mathcal{M}$. As a consequence, each point of $\mathcal{R}$ is dense in $\mathcal{R}$, and similarly for $\mathcal{L}$.

It is a useful exercise to investigate the structure of $\mathcal{N}$ for two-dimensional toroidal space-times, where we identify x with $x + \alpha$ for various values of α.

3 Structures on the Space of Null Geodesics

Although $\mathcal{N}$ as a topological space need not in general be compatible with any manifold structure, we can guarantee that it is in fact a quotient manifold of $N^*\mathcal{M}$ by imposing a standard causal condition.

Theorem 1. *Let $\mathcal{M}$ be strongly causal. Then $\mathcal{N}$, the space of null geodesics of $\mathcal{M}$, inherits a manifold structure from $N^*\mathcal{M}$.*

Proof. If $\mathcal{M}$ is strongly causal, then every point in $\mathcal{M}$ has arbitrarily small neighbourhoods which null geodesics intersect in a single connected component. As a consequence, when the geodesics are lifted to the bundle of null

directions over $\mathcal{M}$, this is also true of the lifts. Then the distribution is regular, and so the quotient space is a quotient manifold [15]. □

In general, if $\mathcal{M}$ is n-dimensional, $T^*\mathcal{M}$ is $2n$-dimensional, so that $N^*\mathcal{M}$ has $2n-1$ dimensions, $\mathbb{P}N^*\mathcal{M}$ has $2n-2$, and so $\mathcal{N}$ has $2n-3$ dimensions. In the standard case, $n=4$ and $\mathcal{N}$ is five-dimensional.

Once we can guarantee that $\mathcal{N}$ is a manifold, we can look for other geometric structures on it. In fact, much of the geometry of $T^*\mathcal{M}$ descends to $\mathcal{N}$. The canonical one-form θ on $T^*\mathcal{M}$ is defined at $\alpha \in T^*\mathcal{M}$ by: for $v \in T_\alpha(T^*\mathcal{M})$, $\theta_\alpha(v) = \alpha(\pi_*(v))$. Then the symplectic form ω on $T^*\mathcal{M}$ is defined by $\omega = \mathrm{d}\theta$.

If $\mathcal{M}$ has local coordinates $\{q^i\}$, and $T^*\mathcal{M}$ has associated coordinates $\{q^i, n_i\}$, then $\theta = n_i \mathrm{d}q^i$ and $\omega = \mathrm{d}n_i \wedge \mathrm{d}q^i$.

The canonical form is the annihilator of a field of hyperplanes on $T^*\mathcal{M}$ which is called a contact structure, see Appendix 4 of Arnold's classical mechanics text [16] for an exposition of contact geometry. Although the form itself is not preserved by dilatations, the field of hyperplanes is, and it is also preserved by the geodesic flow. Consequently, one obtains a field of hyperplanes on $\mathcal{N}$, and in fact also a contact structure on $\mathcal{N}$. Indeed, there is a one-form on $\mathcal{N}$ whose pull-back to $N^*\mathcal{M}$ is proportional to θ, and hence determines the same distribution of hyperplanes, and such a form is a contact form for $\mathcal{N}$.

One can also consider the space $\mathcal{N}'$ of scaled null geodesics and in this case use ω to obtain a symplectic structure on $\mathcal{N}'$ [17]; we will not make use of that structure here.

So let γ be a point on a smooth curve in $\mathcal{N}$, and let j be the tangent to that curve at γ. What does it mean for j to lie in the contact hyperplane at γ?

The vector j at γ determines a Jacobi field, J, along Γ in $\mathcal{M}$ (up to a multiple of the tangent to Γ). Denoting the tangent covector to Γ by n, we then see that j lies in the contact hyperplane at γ iff $n(J) = 0$ at some (and hence any) point of Γ. In other words, the vector connecting two infinitesimally separated null geodesics in $\mathcal{N}$ lies in the contact hyperplane if and only if the vector connecting points on the null geodesics as curves in $\mathcal{M}$ is orthogonal to the tangent to those null geodesics. We say that such null geodesics are abreast.

Penrose has given a detailed exposition of the meanings of θ and ω in the context of null congruences in space-time [18]; the interested reader is referred to this work for more detail.

Clearly, a two-dimensional submanifold of $\mathcal{N}$ corresponds to a smooth two-dimensional family of null geodesics in $\mathcal{M}$, i.e. a three-dimensional surface (perhaps with singularities) ruled by null geodesics.

Note, in passing, that such a surface need not be a null hypersurface:

Example 3. The surface in four-dimensional Minkowski space (with the usual coordinates) consisting of all points on the null geodesics with tangent vector $(1, 1, 0, 0)$ through the surface given by $t = z = 0$ is the timelike surface $z = 0$.

Also, recall that a surface whose tangents all lie in the hyperplanes of the contact structure is called a Legendre surface [16]. Then we finally have

Theorem 2. *Let Σ be a two-dimensional submanifold of $\mathcal{N}$. Then iff Σ is a Legendre surface in $\mathcal{N}$, the surface $\tilde{\Sigma}$ in $\mathcal{M}$ ruled by the null geodesics of Σ is hypersurface-orthogonal; i.e. $\tilde{\Sigma}$ is an orthogonal null congruence to its intersection with any spacelike three-surface in $\mathcal{M}$.*

Proof. From the above discussion we see that a vector connecting neighbouring points of Σ lies in the contact hyperplane if and only if the vector connecting points of nearby null generators of $\tilde{\Sigma}$ is orthogonal to the tangents to the null generators. Hence the tangent vector to the intersection of a spacelike three-surface with $\tilde{\Sigma}$ is orthogonal to the tangent to any null generator of $\tilde{\Sigma}$, i.e. $\tilde{\Sigma}$ is an orthogonal null congruence to this intersection. □

In particular, if $p \in \mathcal{M}$, then we can find the subset of $\mathcal{N}$ consisting of all null geodesics through p. This subset is the image of the S^2 fibre over p in the bundle of null directions over $\mathcal{M}$, and is itself a smooth S^2 in $\mathcal{N}$, which we denote P; the S^2 in $\mathcal{N}$ corresponding to a point of $\mathcal{M}$ is called the sky of that point. Every sky is a Legendre surface in $\mathcal{N}$; but a Legendre surface need not be a sky.

Even though the space of null geodesics of a strongly causal space-time is naturally a manifold, it may still have pathologies: in particular, it may fail to be Hausdorff.

Example 4. Consider Minkowski space with the usual coordinates (t, x, y, z), and let its cotangent bundle have coordinates $(t, x, y, z, p_t, p_x, p_y, p_z)$. Let $\mathcal{M}$ be Minkowski space minus the origin. Then the sequence of null geodesics given by the covectors $(0, 1/n, 0, 0, 1, 1, 0, 0)$ has as limit points both the null geodesic determined by the covector $(1, 1, 0, 0, 1, 1, 0, 0)$ and that determined by $(-1, -1, 0, 0, 1, 1, 0, 0)$. (In Minkowski space these covectors determine the same null geodesic, which passes through the origin.)

Such a pathology cannot arise in the case where $\mathcal{M}$ is globally hyperbolic. In fact, if $\mathcal{S}$ is a Cauchy surface for $\mathcal{M}$, then $\mathcal{S}$ inherits a Riemannian structure from g, and $\mathcal{N}$ is diffeomorphic to the tangent unit sphere bundle to $\mathcal{S}$ with this Riemannian metric; furthermore, the contact structure on $\mathcal{N}$ agrees with the natural one on the tangent unit sphere bundle of $\mathcal{S}$. In this case, being the tangent unit sphere bundle of a Hausdorff (Riemannian) manifold, $\mathcal{N}$ is automatically Hausdorff.

This immediately tells us that any space-time whose space of null geodesics is not Hausdorff cannot be globally hyperbolic: so in particular, neither

Minkowski space with a point removed, nor the impulsive gravitational plane wave space-time [19] can be globally hyperbolic.

Global hyperbolicity is a sufficient, but not a necessary condition for $\mathcal{N}$ to be Hausdorff. For example, the region of Minkowski space given by $x^2 + y^2 + z^2 < 1$ has a Hausdorff space of null geodesics, but is not globally hyperbolic.

4 Insight into Space-Time

A natural question to ask is how the causal structure of $\mathcal{M}$ is reflected in $\mathcal{N}$.

We will use the convention of taking lower case Latin letters to refer to points of $\mathcal{M}$, and the corresponding upper case letter to represent the sky of a point; also, a lower case Greek letter will represent a point of $\mathcal{N}$, and the corresponding upper case letter the corresponding null geodesic curve in $\mathcal{M}$ (or, more precisely, its image).

The situation is simplest in Minkowski space, where $\mathcal{N}$ is projective null twistor space, excluding the line at infinity, $\mathbb{PN}^I$, which has been extensively studied from the point of view of projective geometry [18]. In this case $\mathcal{N}$ has a great deal of extra structure; in particular, it is naturally a real submanifold of $\mathbb{CP}^3$, and skies are characterised as the holomorphic surfaces with topology S^2.

Then if $x_1, x_2 \in \mathcal{M}$, x_1 and x_2 lie on a common null geodesic (i.e. are null separated) iff $X_1 \cap X_2$ is non-empty; and dually, if $\gamma_1, \gamma_2 \in \mathcal{N}$, then γ_1 and γ_2 lie on a common sky iff $\Gamma_1 \cap \Gamma_2$ is non-empty.

Null separation, then, is neatly described in this picture, and there is a natural duality between points and null geodesics in space-time and points and skies in the space of null geodesics.

But it is not only null separation which can be given an elegant characterization. $\mathbb{PN}^I$ is topologically $\mathbb{R}^3 \times S^2$, and it is possible to define a linking number L for two skies in $\mathbb{PN}^I$. This linking number may be computed in Minkowski space. Given p and q, let $\mathcal{S}$ be a surface of constant time containing p; then the light cone of q, $N(q)$, intersects $\mathcal{S}$ in an S^2 in general. The linking number of Q round P is simply the winding number of $N(q) \cap \mathcal{S}$ round p in $\mathcal{S}$. By the appropriate choice of orientation and sign convention [20], we have $p \in I^{\pm}(q)$ iff $L(P, Q) = \pm 1$.

In more general, curved, space-times, we lose much of the structure of twistor theory: in compensation, new phenomena arise.

As before, we observe that a Jacobi field, J, along the null geodesic Γ arising from a one-parameter family of null geodesics in $\mathcal{M}$ determines a tangent vector, j, at $\gamma \in \mathcal{N}$, and vice versa. Furthermore, J is tangent to Γ at p if and only if $j \in T_\gamma P$. Recall that two points, p and q are conjugate [8] along Γ if and only if there is a non-trivial Jacobi field along Γ which is tangent to Γ at p and q. Because of this, we have:

Theorem 3. *Let $x, y \in \mathcal{M}$ lie on the null geodesic Γ. Then $\gamma \in X \cap Y$. Furthermore, X and Y interesect transversally in γ unless x is conjugate to y along Γ, and in this case the dimension of $T_\gamma X \cap T_\gamma Y$ is the number of linearly independent Jacobi fields along Γ vanishing at both x and y.*

The property of Minkowski space, that points p and q are chronologically related iff P and Q are linked in $\mathbb{PN}^I$, fails in general, even in globally hyperbolic space-times. It still holds if p and q are close together, in the following sense:

Theorem 4. *Let $\mathcal{M}$ be strongly causal, and for $p \in \mathcal{M}$ let U be a causally convex neighbourhood of p (so U is geodesically convex, and causal curves intersect U in a single connected component). Then denoting the space of null geodesics of U by $\mathcal{N}(U)$, chronological relations of points in U are encoded in the linking of their skies in $\mathcal{N}(U)$ in just the same way as in Minkowski space.*

In other words, linking of skies still encodes chronological relations locally; but globally it need not. If $N^+(p)$ has self-intersections, there can be points $q \in I^+(p)$ such that $\mathrm{Link}(P, Q) = 0$. In fact, if we use the equivalence of $\mathcal{N}$ with the unit tangent sphere bundle to $\mathcal{S}$ in a globally hyperbolic space-time with Cauchy surface $\mathcal{S}$, one can have points p and q with $q \in I^+(p)$ but such that P and Q can be simultanously deformed to tangent spheres [21]. If $\mathcal{S}$ is a Cauchy surface containing p, this will occur when $N(q) \cap \mathcal{S}$ has winding number zero round p.

So there is no sense in which topological linking encodes causal relations in general.

At least, not in four (or more) space-time dimensions. If we consider the case of three space-time dimensions, the situation is rather different. As is well known, topology in three dimensions has special properties; in particular, it is possible to have two S^1s embedded in S^3 in such a way that they have linking number 0, but are nevertheless non-trivially linked [22]. One particular example of such a link is the Whitehead link; and the example alluded to above reduces precisely to the Whitehead link if we reduce the number of spatial dimensions by one.

Conjecture 1. Let $\mathcal{M}$ be a globally hyperbolic space-time with two spatial dimensions, and Cauchy surface $\mathcal{S}$ diffeomorphic to $\mathbb{R}^2$, so that $\mathcal{N}$ is diffeomorphic to $\mathbb{R}^2 \times S^1$. Then points p and q are causally related if and only if their skies cannot be simultaneously deformed to unit tangent spheres of $\mathcal{S}$.

A proof of this conjecture in a non-trivial class of space-times has been found [23], although the general case remains elusive.

To return to the more physically interesting case of four-dimensional space-time, we see that the problem is caused by light cones developing self-intersections. In fact, the apparatus we now have available provides a

powerful tool for investigating just how the light cone can develop in a general space-time. This provides information on the types of caustic that can arise due to gravitational lensing.

First, we recall that, by Theorem 2, a Legendre surface Σ in $\mathcal{N}$ corresponds to a hypersurface-orthogonal null hypersurface $\tilde{\Sigma}$ in $\mathcal{M}$; but saying that $\tilde{\Sigma}$ is hypersurface-orthogonal is just saying that it is the wave front obtained by instantaneously lighting up an initial space-like two-surface and tracing out the resulting light rays. So Legendre surfaces of $\mathcal{M}$ correspond to wave fronts in $\mathcal{M}$. Furthermore, if $\mathcal{S}$ is a Cauchy surface of $\mathcal{M}$, and we consider $\mathcal{N}$ as the unit tangent sphere bundle to $\mathcal{S}$, then the natural projection of Σ to $\mathcal{S}$ gives the intersection of $\tilde{\Sigma}$ with $\mathcal{S}$; this projection is a Legendre map. We can then deduce from the properties of such mappings presented by Arnold [16] that the only singularities which are present in the intersection of a wave front with a Cauchy surface and stable under small perturbations are those of type A_2 or D_4. In particular, these are the only stable singularities which appear on a light cone at a given instant.

One can alternatively carry out an analysis in terms of the full structure of the cotangent bundle of space-time [24]. Such an analysis allows for the investigation of other properties of wave front evolution, but at the expense of requiring far more technical machinery. An extensive review of gravitational lensing and wave front evolution has now been provided by Perlick [25].

5 Recovering Space-Time

We have considered how points of $\mathcal{M}$ are represented in $\mathcal{N}$ as skies, and seen how this gives a new approach to considerations of the causal structure of $\mathcal{M}$. A natural question to ask is whether we can use such a setting to define a space-time as a family of skies in a suitable five-manifold; and in such a setting, to look for characterizations of a space-time being conformal to an Einstein vacuum, or to a space-time satisfying a standard energy condition.

Unfortunately, such a characterization is not (yet) available. Indeed, there may be problems even with recovering the original space-time if we know all the skies in $\mathcal{N}$.

Example 5. Let $\mathcal{M}$ be the Einstein static universe. Then the space of null geodesics is precisely that of compactified and identified Minkowski space [18], and all the skies of points in $\mathcal{M}$ are skies of points of compactified, identified Minkowski space.

The phenomenon at the heart of this problem is that the null cone of a point in the Einstein static universe converges back to a point again; so distinct points can have the same sky. Clearly, this is a bad thing from the point of view of regarding a space-time as arising from a set of skies.

In fact, even if the null cone does not converge back exactly to a point, there may still be a problem. For suppose p is a point of $\mathcal{M}$, and p_n is a

sequence of points which remain strictly outside some neighbourhood K of p, but have the property that if U is a neighbourhood of p then for n sufficiently large, all the null geodesics through p_n pass through U. We call a space-time exhibiting this behaviour a *refocussing* space-time.

Then in $\mathcal{N}$, no matter how small a neighbourhood V of P we choose, there will be infinitely many P_n lying inside V. In this case, even if all the points of $\mathcal{M}$ have distinct skies, $\mathcal{N}$ cannot provide $\mathcal{M}$ with the correct topology.

Fortunately, this phenomenon cannot occur in a large class of space-times of interest.

Theorem 5. *Let $\mathcal{M}$ be globally hyperbolic, with non-compact Cauchy surface $\mathcal{S}$. Then $\mathcal{M}$ cannot be refocussing.*

Proof. In brief outline: We can suppose without loss of generality that each $p_n \in I^+(p)$. Now, if the entire light cone of p_n focusses back into a small neighbourhood of p, then all null geodesics through p_n must meet a conjugate point within some finite time. In this case, it follows that all of $\mathcal{S}$ lies in $J^-(p_n)$, which is impossible because $J^-(p_n) \cap \mathcal{S}$ must always be compact. □

If we are given the set of all skies in the space of null geodesics, $\mathcal{N}$, of a globally hyperbolic space-time, $\mathcal{M}$, with non-compact Cauchysurface, $\mathcal{S}$, then we can reconstruct the original space-time up to a conformal factor.

First, we have $\mathcal{M}$ simply as the point set of all skies in $\mathcal{N}$.

Next, if P is a sky in $\mathcal{N}$, we take a neighbourhood U of P, sufficiently small, that any two skies in P intersect transversally. Then the set of all skies lying in U give a neighbourhood, V, of p. Neighbourhoods constructed in such a way give a basis for the topology on $\mathcal{M}$.

Recovering the differentiable structure of this neighbourhood is a little more involved. First, we construct the Grassmannian bundle of two-planes over U; then we lift each sky in U to this bundle by lifting $\gamma \in Q$ to $T_\gamma Q$. This gives a six-dimensional submanifold $\tilde{U}$ of our Grassmannian manifold which is (because of the smoothness of the geodesic flow and the absence of non-transversal intersections) diffeomorphic to the bundle of null directions over V. Each sky lifts to the fibre over a point of V, and so taking the quotient manifold of lifted skies (fibres here are compact, so the distribution is automatically regular) gives us back V as a differentiable manifold.

It is now a simple matter to find the metric, g, up to a conformal factor. Given a point $\gamma \in U$, we obtain a curve in $\tilde{U}$ given by the tangent planes to all skies in U containing γ. This projects to a curve in V under the quotient above, namely Γ. But once we know all the null geodesics in V, we have the null cone at each point, and as is well known [8] this determines the metric up to a conformal factor.

We can then recover $\mathcal{M}$ by constructing a neighbourhood of each point by this means, and using the overlap maps induced by the intersections of neighbourhoods of skies in $\mathcal{N}$.

Alternatively, we can make use of an alternative topology on $\mathcal{M}$ which was observed by Hawking et al. [26] to capture all of the relevant structure of $\mathcal{M}$.

Again, we begin with a $\mathcal{M}$ as the point set of skies in $\mathcal{N}$, but no other structure, and consider a small neighbourhood U of the sky P in $\mathcal{M}$. This time, we take as a neighbourhood of $p \in \mathcal{M}$ all those skies in U which are nontrivially linked to P. This gives a basis for the topology on $\mathcal{M}$ generated by neighbourhoods of the form $(V \cap I(p)) \cap \{p\}$, where V is a neighbourhood of p in the original manifold topology, and I is determined by the chronology relation of the original Lorentz metric on $\mathcal{M}$. This topology, called the path topology, uniquely determines the causal, differential, and conformal structure of $\mathcal{M}$.

Thus, at least in the case of non-refocussing space-times, it is possible to regard space-time as given by a structure (namely the set of skies) in the space of null geodesics. In the final section we will consider an approach to providing null geodesics with endpoints, in an attempt to provide space-time with a causal boundary.

6 A (New?) Causal Boundary

Let $\mathcal{M}$ be a strongly causal space-time, and let γ be a null geodesic of $\mathcal{M}$. How can we attach a future endpoint to Γ? The idea behind what will be done here is to find all null geodesics which focus at the same point at infinity, and treat this set of null geodesics as the light cone of the (common) future endpoint of these null geodesics.

To this end, we let s be an affine parameter for Γ, so that $p(s)$ traces out Γ as s ranges from $-\infty$ to ∞. (By an appropriate choice of conformal factor, we can assume that all null geodesics have affine parameters with this range of values [27].) Then as s increases, $T_\gamma P(s)$ traces out a curve in the Grassmannian manifold of two-dimensional subspaces of $T_\gamma \mathcal{N}$. Since Grassmannian manifolds are compact, this curve has a limit point as $s \to \infty$.

This limiting two-plane is supposed to be the tangent plane to the sky of the future endpoint of Γ. However, it need not be unique. One would expect that $T_\gamma P(s)$ would settle down if curvature decayed along Γ; ascertaining appropriate conditions for uniqueness is the subject of current work. Thus uniqueness should be related to the asymptotic flatness of the (perhaps conformally rescaled) space-time.

Then at worst, we have some subset - denote it by B_γ - of $T_\gamma \mathcal{N}$ for each $\gamma \in \mathcal{N}$. In general, this need not be a distribution; its dimension may vary from point to point, and it need not be continuous. Now regard γ_1 and γ_2 as equivalent if they can be connected by a curve whose tangent everywhere lies in some B_γ. By definition, null geodesics which are equivalent under this relationship focus to a common future endpoint.

Then we obtain a topological space of future endpoints to the null geodesics of $\mathcal{M}$ by taking the quotient space under this relation: call this topological space B^+ (for future boundary). An equivalence class (point at infinity) will be denoted by B when we are thinking of it as a subset of $\mathcal{N}$, and by b when we think of it as a point of the boundary of $\mathcal{M}$. What we still lack is a topology for $\mathcal{M} \cup B^+$; so how do we decide if a sequence of points p_n in $\mathcal{M}$ converges to a point, b, of B^+?

We require two conditions on the sequence of points p_n to say that $p_n \to b$. First, p_n should eventually leave any compact set in $\mathcal{M}$; and second, the light cone of p_n should approach $N^-(b)$ (which is, by definition, the set of null geodesics defining b). To make this latter condition precise, we require that there exists $\gamma \in B$ and $\gamma_n \in P_n$ such that $\gamma_n \to \gamma$, and the limit set of T_{γ_n} lies in B_γ as $n \to \infty$.

This certainly works well in certain simple cases.

If $\mathcal{M}$ is the Einstein static cylinder, then all null geodesics share a single future endpoint, which lies to the future of every point of $\mathcal{M}$.

More generally, if $\mathcal{M}$ can be conformally embedded into a strongly causal space-time as a subspace with compact closure, then each null geodesic will acquire a future endpoint. Furthermore, each of the equivalence classes defined above will be a subset of the sky of such a future endpoint. If we further require that $\mathcal{M}$ be globally hyperbolic, then the equivalence classes will coincide with the skies of endpoints, and the boundary points are precisely those of the usual causal boundary [10].

So this attempt at constructing a boundary is well behaved in certain simple cases where we have a good idea of what the boundary "ought" to be. Furthermore, properties of space-time which complicate the intuitive notion of boundary also complicate this construction.

Note that one can similarly attempt to add a past endpoint to each null geodesic, and thereby construct a past boundary, B^-. However, the usual problem of identification of appropriate points in the past and future boundary remains.

In conclusion, I will list some of the questions which arise in the context of this proposal:

1. How is this boundary related to the Geroch, Kronheimer and Penrose boundary?
2. For example, if null geodesics lie in the same equivalence class, need they have the same chronological past in general?
3. How is this boundary related to the Geroch boundary [28]? Is his equivalence relationship strictly weaker than this one, vice versa, or are they incomparable?
4. Does good/bad behaviour of the limiting "distribution" match to good/bad asymptotic properties of the original space-time?
5. In particular, can we gain any insight into space-time in the vicinity of a strong curvature singularity from this point of view?

References

1. C. Will: *Theory and Experiment in Gravitational Physics* (Cambridge University Press, Cambridge 1993) 35
2. `http://www.gravityprobeb.com/` 35
3. `http://www.ligo.caltech.edu/` 35
4. K. Gödel: An example of a new type of cosmological solution of Einstein's field equations of gravitation. Rev. Mod. Phys. **21**, 447–450 (1943) 36
5. J. Earman: *Bangs, Crunches, Whimpers and Shrieks: Singularities and Acausalities in Relativistic Spacetimes* (Oxford University Press, Oxford 1995) 36
6. R. Penrose: Singularities and time asymmetry. In: *General Relativity: An Einstein Centenary Survey*, ed by S.W. Hawking, W. Israel (Cambridge University Press, Cambridge 1979) 36
7. R. Penrose: *Techniques of Differential Topology in Relativity*, Regional Conference Series in Applied Math. **7** (Society for Industrial and Applied Mathematics, Philadelphia 1972) 37
8. S.W. Hawking, G.F.R. Ellis: *The Large Scale Structure of Space-Time* (Cambridge University Press, Cambridge 1973) 37, 43, 46
9. J.K. Beem, P.E. Ehrlich: *Global Lorentzian Geometry* (Marcel Dekker, New York 1981) 37
10. R.P. Geroch, E.H. Kronheimer, R. Penrose: Ideal points of space-time. Proc. Roy. Soc. London A **327**, 545–567 (1972) 37, 48
11. R.D. Sorkin: Forks in the road, on the way to quantum gravity. Int. J. Theor. Phys. **36**, 2759–2781 (1997) 37
12. J.F. Cariñena, C. López: Symplectic structure on the set of geodesics of a Riemannian manifold. Int. J. Modern Phys. A **6**, 431–444 (1991) 38
13. J.K. Beem, R.J. Low, P.E. Parker: Spaces of geodesics: Products, coverings, connectedness. Geometriae Dedicata **59**, 51–64 (1996) 38
14. B. Carter: Causal structure in space-time. Gen. Rel. Grav. **1**, 349–391 (1971) 38
15. F. Brickell, R.S. Clark: *Differentiable Manifolds: An Introduction* (Van Nostrand Reinhold, London 1970) 39, 41
16. V.I. Arnold: *Mathematical Methods of Classical Mechanics*, 2nd edn (Springer, New York 1991) 41, 42, 45
17. R. Penrose: On the nature of quantum geometry. In: *Magic Without Magic: J. A. Wheeler Festschrift*, ed by J. R. Klauder (Freeman, New York 1972) 41
18. R. Penrose, W. Rindler: Spinors and space-time. Vol 2: Spinor and Twistor Methods in space-time Geometry (Cambridge University Press, Cambridge 1986) 41, 43, 45
19. R. Penrose: A remarkable property of plane waves in general relativity. Rev. Mod. Phys. **37**, 215–220 (1965) 43
20. R.J. Low: Twistor linking and causal relations. Class. Quantum Grav. **7**, 177–187 (1990) 43
21. R.J. Low: Causal relations and spaces of null geodesics. D. Phil. Thesis, Mathematical Institute, Oxford University (1988) 44
22. D. Rolfsen: *Knots and Links* (AMS, Chelsea 2003) 44
23. J. Natário, K. P. Tod: Linking, Legendrian linking and causality. Proc. London Math. Soc. **88**, 251–272 (2004) 44
24. W. Hasse, M. Kriele, V. Perlick: Caustics of wavefronts in general relativity. Class. Quantum Grav. **13**, 1161–1182 (1996) 45

25. V. Perlick: *Gravitational Lensing from a Spacetime Perspective*, Living Rev. Relativity **7** (2004), 9. `http://www.livingreviews.org/lrr-2004-9` 45
26. S.W. Hawking, A.R. King, P.T. McCarthy: A new topology for space-time which incorporates the causal, differential and conformal structures. J. Math. Phys. **17**, 174–181 (1976) 47
27. J.K. Beem: Conformal changes and geodesic completeness. Comm. Math. Phys. **49**, 179–186 (1976) 47
28. R. Geroch: Local characterization of singularities in general relativity. J. Math. Phys. **9**, 450–465 (1967) 48

Some Variational Problems in Semi-Riemannian Geometry

Antonio Masiello

Dipartimento di Matematica, Politecnico di Bari, Via Amendola 126/B, Bari, 70125, Italy
masiello@poliba.it

Abstract. In this contribution we are concerned with global properties of geodesics on semi-Riemannian manifolds obtained by studying the variational properties of the action functional. Applications to physically meaningful spacetimes in General Relativity will be presented.

1 Introduction

We consider a semi-Riemannian manifold $(\mathcal{M}, g)$, where $\mathcal{M}$ is a smooth, connected, finite dimensional differentiable manifold and g is a metric tensor on $\mathcal{M}$. For any $z \in \mathcal{M}$, the tensor g defines a bilinear form $g(z)$ on the tangent space $T_z\mathcal{M}$ at z to $\mathcal{M}$ such that $g(z)$ is symmetric and nondegenerate. The number of negative eigenvalues of the bilinear form $g(z)$ does not depend on z; this number is called the *index* of the metric g and it is denoted by $\nu(g)$. The semi-Riemannian manifold $(\mathcal{M}, g)$ is called *Riemannian* if $\nu(g) = 0$ and it is called *Lorentzian* if $\nu(g) = 1$. We refer to the books [4, 46, 55, 68] for the basic properties of semi-Riemannian manifolds and physical properties of spacetimes.

A smooth curve $\gamma : I \longrightarrow \mathcal{M}$, where I is an interval of the real line $\mathbf{R}$, is called a *geodesic* if

$$\nabla_s \dot{\gamma} = 0 \,, \tag{1}$$

where ∇_s denotes the covariant derivative along γ induced by the Levi-Civita connection of g and $\dot{\gamma}$ is the tangent vector field along γ. In local coordinates, equation (1) reduces to the system of $\dim(\mathcal{M})$ nonlinear second order differential equations

$$\ddot{\gamma}^k + \Gamma^k_{ij}(\gamma)\dot{\gamma}^i\dot{\gamma}^j = 0 \,, \qquad k = 1, \ldots, \dim(\mathcal{M}) \,,$$

where the Γ^k_{ij} are the Christoffel symbols of the metric g. In this paper we shall focus our attention on the problem of the existence of one or multiple geodesics joining two arbitrary points of a semi-Riemannian manifold.

Definition 1. *A semi-Riemannian manifold $(\mathcal{M}, g)$ is said to be* geodesically connected *if any pair of points p and q of $\mathcal{M}$ is joined by a geodesic for the metric g.*

A. Masiello: *Some Variational Problems in Semi-Riemannian Geometry*, Lect. Notes Phys. **692**, 51–77 (2006)
www.springerlink.com

For Riemannian manifolds, the problem of geodesic connectedness is essentially solved. Indeed, as a consequence of the Hopf–Rinow Theorem, any metrically complete (or, equivalently, geodesically complete) Riemannian manifold is geodesically connected. Moreover, any pair of points is joined by infinitely many geodesics if the manifold $\mathcal{M}$ is noncontractible.

The problem of geodesic connectedness is much more delicate for semi-Riemannian manifolds, and actually there are only few intrinsic results. Moreover, many meaningful counter-examples to the geodesic connectedness of a Lorentzian manifold are known. We mention the remarkable and in some way surprising result by Calabi and Markus who showed in [12], see also [55, p. 248], that for $\dim(\mathcal{M}) \geq 3$ a *Lorentzian space-form with positive curvature is geodesically connected if and only if it is not time-oriented*. For instance the *de Sitter spacetime* is not geodesically connected. Counter-examples to the geodesic connectedness are also contained in the following classes of Lorentzian manifolds, see [55]:

- *compact Lorentzian manifolds*: the Clifton–Pohl torus;
- *geodesically complete Lorentzian manifolds*: the de Sitter and the anti-de Sitter spacetimes;
- *globally hyperbolic Lorentzian manifolds*: again the de Sitter spacetime.

The first global existence results on geodesics on Lorentzian manifolds (locally the geodesic connectedness always holds in a convex neighborhood of a point) concerned timelike and lightlike geodesics, because of their physical meaning. The Avez–Seifert Theorem states that any pair of causally related points in a globally hyperbolic spacetime is joined by a causal geodesic [4, 55]. Moreover, due to the application to the multiple image effect in gravitational lensing, the variational theory of light rays and the extension to General Relativity of the classical Fermat principle has been very intensively investigated. In this direction see the paper by Uhlenbeck [67], the papers [27, 34, 41, 39] and the recent living review by Perlick [59]. In this paper we present some results on the geodesical connectedness of some classes of semi-Riemannian manifolds obtained in the last years, using variational methods for strongly indefinite functionals.

It is well known that the geodesics joining two points on a semi-Riemannian manifold satisfy a variational principle. Let $(\mathcal{M}, g)$ be a semi-Riemannian manifold; then the geodesics joining two points p and q on $\mathcal{M}$ are the stationary points of the action integral

$$f(z) = \int_0^1 g(z(s))[\dot{z}(s), \dot{z}(s)]ds \tag{2}$$

defined on the infinite dimensional Sobolev manifold $\Omega^{1,2}(p, q; \mathcal{M})$ of the curves $z(s) : [0, 1] \longrightarrow \mathcal{M}$ such that $z(0) = p$, $z(1) = q$, z is continuous and its derivative $\dot{z}$ is square-integrable with respect to some Riemannian metric on $\mathcal{M}$ (this definition does not depend on the choice of the Riemannian

metric). It is well known that the space $\Omega^{1,2}(p,q;\mathcal{M})$ is equipped with the structure of an infinite dimensional manifold modelled on the Sobolev–Hilbert space $H^{1,2}([0,1],\mathbf{R}^n)$ of absolutely continuous curves in $\mathbf{R}^n$, $n=\dim(\mathcal{M})$, having square-integrable derivative. If $z\in\Omega^{1,2}(p,q;\mathcal{M})$, the tangent space $T_z\Omega^{1,2}(p,q;\mathcal{M})$ at z is given by

$$T_z\Omega^{1,2}(p,q;\mathcal{M})=\{\zeta\in\Omega^{1,2}((p,0),(q,0);T\mathcal{M}):\pi\circ\zeta=z\}\,, \tag{3}$$

where $T\mathcal{M}$ denotes the tangent bundle of $\mathcal{M}$ and $\pi\colon T\mathcal{M}\to\mathcal{M}$ is the bundle projection. In other words $T_z\Omega^{1,2}(p,q;\mathcal{M})$ consists of the vector fields ζ along z of class $H^{1,2}$ that vanish at the end-points.

The study of the global properties of geodesics on a Riemannian manifold, such as existence and multiplicity and also the existence of a closed geodesic, and the relations between the set of such geodesics and the topology of the underlying manifold $\mathcal{M}$, have played a central role in the XX century in the development of what is called now the *Calculus of Variations in the Large* and in particular in the *Critical Point Theory*, the study of the critical points of a functional which include not only global or local minima (or maxima), as in the classical Calculus of Variations, but also saddle points. In particular the min-max Ljusternik–Schnirelmann Theory and Morse Theory have been successfully applied to the action integral in Riemannian geometry, and the problems of the existence and the multiplicity of closed geodesics on compact Riemannian manifolds have been recently solved, see [48]. Moreover many relations between curvature and topology of a complete Riemannian manifold can be obtained using Morse Theory. In a certain sense the action integral of a complete Riemannian manifold is the functional where global variational methods work best.

The situation drastically changes if we pass to semi-Riemannian manifolds, in particular to Lorentzian manifolds. Indeed, the action integral (2) for semi-Riemannian metrics is one of the worst functionals to apply variational methods. It is an example of a *strongly indefinite functional*, characterized by the following problems:

- $f(z)$ is unbounded both from below and from above, so a critical point of f cannot be found by a minimization (or a maximization) argument as for the action integral in Riemannian geometry.
- The Morse index (respectively coindex) of any stationary point z of f is equal to $+\infty$. This means that the second derivative $f''(z)$ at the critical point z is negative (respectively positive) definite on an infinite dimensional subspace of the tangent space $T_z\Omega^{1,2}(p,q;\mathcal{M})$. Thus any critical point of f in the semi-Riemannian case is an infinite dimensional saddle point. This fact makes more difficult the search of critical points by the methods of the gradient flow, see Sect. 2.
- The action functional f does not satisfy the Palais–Smale condition, see Sect. 2.

Critical Point Theory for strongly indefinite functionals has been the object of several deep studies in the last twenty-five years, and it has many applications to the study of nonlinear differential equations. We refer to the books [1, 19, 52, 64] for abstract results on strongly indefinite functionals and applications to Hamiltonian systems, nonlinear hyperbolic equations, wave maps and symplectic geometry, and to [51] for an introduction to variational methods oriented towards the study of geodesics on Lorentzian manifolds.

In this paper we shall present a review of critical point theorems and some applications to geodesics on semi-Riemannian manifolds. The paper is organized as follows. In Sect. 2 a review of variational methods is presented. In Sect. 3 the abstract results are applied to geodesics on Riemannian manifolds and in Sect. 4 to stationary Lorentzian manifolds, the two classes of manifolds for which the geodesic problem is not strongly indefinite or can be reduced to a not strongly indefinite one. In Sect. 5 the geodesic problem for orthogonally splitting Lorentzian manifolds is considered. In Sect. 6 results for physically relevant spacetimes of General Relativity are stated. Finally in Sect. 7 other directions in the study of variational properties of semi-Riemannian manifolds are presented.

2 A Review of Variational Methods

We present in this section the main results of Critical Point Theory, for the proofs see [19, 51, 52, 64].

Let (X, h) be a smooth (C^∞), possibly infinite dimensional Riemannian manifold and $f : X \to \mathbf{R}$ a C^1 functional. Then a point $x \in X$ is said to be a *critical point* of f if $f'(x) = 0$. A number $c \in \mathbf{R}$ is called a *critical value* if there exists a critical point x of f such that $f(x) = c$, otherwise c is called a *regular value*. Let x be a critical point of f, denote by T_xX the tangent space at x to X and assume that f is of class C^2. Then the Hessian $H_f(x) : T_xX \times T_xX \to \mathbf{R}$ at x is defined in the following way. For any $\xi \in T_xX$ we set

$$H_f(x)\,[\xi, \xi] = \left(\frac{d^2 f(\eta(s))}{ds^2} \right)_{s=0}$$

(where $\eta :]-\varepsilon, \varepsilon[\to X$ is a smooth curve such that $\eta\,(0) = x, \dot\eta(0) = \xi$) and then we extend $H_f(x)$ by polarization to any pair of tangent vectors. The *Morse index* $m(x, f)$ of a critical point x of f is the maximal dimension of a subspace of T_xX where $H_f(x)$ is negative definite. The *augmented Morse index* $m^*(x, f)$ is defined as $m^*(x, f) = m(x, f) + \dim\big(\ker H_f(x)\big)$, where

$$\ker H_f(x) = \{\xi \in T_xX : H_f(x)[\xi, \xi'] = 0, \forall \xi' \in T_xX\}\,.$$

Clearly, the Morse index and the augmented Morse index may be equal to $+\infty$ if X is an infinite dimensional Riemannian manifold. The critical point x is said to be *nondegenerate* if the linear operator $f''(x) : T_xX \to T_xX$

induced by $H_f(x)$ on T_xX, equipped with the Hilbert space structure $h(x)$, is an isomorphism. The functional f is said to be a *Morse function* if it is of class C^2 on the manifold (X, h) and all its critical points are nondegenerate.

In order to study the variational properties of geodesics as critical points of the action integral f defined in (2), it is essential to evaluate the Hessian $H_f(z)$ at a geodesic z. Let $(\mathcal{M}, g)$ be a semi-Riemannian manifold, fix two points p and q in $\mathcal{M}$ and consider the infinite dimensional manifold $\Omega^{1,2}(p, q; \mathcal{M})$. The tangent space $T_z\Omega^{1,2}(p, q; \mathcal{M})$ is given by (3). We can equip the manifold $\Omega^{1,2}(p, q; \mathcal{M})$ with the structure of an infinite dimensional Riemannian manifold in the following way. Choose a Riemannian metric g_0 on $\mathcal{M}$ and denote by ∇^0 the Levi–Civita connection for the metric g_0; then a Riemannian metric h_0 on $\Omega^{1,2}(p, q; \mathcal{M})$, depending on g_0, is defined as follows: for any $z \in \Omega^{1,2}(p, q; \mathcal{M})$ and for any pair of tangent vector fields ζ, $\zeta' \in T_z\Omega^{1,2}(p, q; \mathcal{M})$, the bilinear form $h_0(z)$ is defined by setting

$$h_0(z)[\zeta, \zeta'] = \int_0^1 g_0(z)[\nabla_s^0\zeta, \nabla_s^0\zeta']\, ds\ .$$

It is well known that all the metrics h_0 are equivalent, independently of the Riemannian metric g_0 on the physical manifold $\mathcal{M}$, so that they induce the same topology on the manifold $\Omega^{1,2}(p, q; \mathcal{M})$. Consider now the action integral

$$f(z) = \int_0^1 g(z)[\dot{z}, \dot{z}]\, ds\ .$$

It is well known that the functional f is smooth on $\Omega^{1,2}(p, q; \mathcal{M})$ and, for any $z \in \Omega^{1,2}(p, q; \mathcal{M})$ and $\zeta \in T_z\Omega^{1,2}(p, q; \mathcal{M})$, the first variation $f'(z)[\zeta]$ is given by

$$f'(z)[\zeta] = \int_0^1 g(z)[\dot{z}, \nabla_s\zeta]\, ds\ ,$$

where ∇ is the Levi–Civita connection with respect to the semi-Riemannian metric g. A curve $z \in \Omega^{1,2}(p, q; \mathcal{M})$ is a critical point for f if and only the first variations $f'(z)[\zeta]$ are zero for any admissible direction $\zeta \in T_z\Omega^{1,2}(p, q; \mathcal{M})$. An integration by parts and a classical boot-strap argument show that a curve z is a critical point of the action integral f on $\Omega^{1,2}(p, q; \mathcal{M})$ if and only if z is a geodesic for the metric g satisfying $z(0) = p$ and $z(1) = q$.

Now, let z be a geodesic joining p and q; then the Hessian $H_f(z)$ of the action integral at the critical point z is given by (see for instance [51]):

$$H_f(z)[\zeta, \zeta'] = \int_0^1 g(z)[\nabla_s\zeta, \nabla_s\zeta']\, ds - \int_0^1 g(z)[R(\zeta, \dot{z})\dot{z}, \zeta']\, ds\ , \qquad (4)$$

for any ζ,$\zeta' \in T_z\Omega^{1,2}(p, q; \mathcal{M})$, where $R(\cdot, \cdot)\cdot$ denotes the curvature tensor for the metric g.

Let $f''(z)\colon T_z\Omega^{1,2}(p, q; \mathcal{M}) \times T_z\Omega^{1,2}(p, q; \mathcal{M}) \to \mathbf{R}$ be the linear operator on $T_z\Omega^{1,2}(p, q; \mathcal{M})$ (equipped with the Hilbert space structure induced by

$h_0(z)$) associated with the Hessian $H_f(z)$. The operator $f''(z)$ is a symmetric Fredholm operator of index 0, because it is a compact perturbation of an isomorphism on $T_z\Omega^{1,2}(p,q;\mathcal{M})$. But equation (4) clearly shows how the index $\nu(g)$ influences the spectral properties of $f''(z)$. Indeed, if g is a Riemannian metric, the linear operator $f''(z)$ on $T_z\Omega^{1,2}(p,q)$ is a compact perturbation of a positive definite bilinear form. On the other hand, if $\nu(g) > 0$, then $f''(z)$ is still a Fredholm operator, but now it is a compact perturbation of a nondegenerate symmetric bilinear form which is both negative definite and positive definite on some infinite dimensional subspace of $T_z\Omega^{1,2}(p,q;\mathcal{M})$. Thus the Morse index $m(z,f)$ of any geodesic is finite for Riemannian metrics, but it is equal to $+\infty$ if $\nu(g) > 0$. Any geodesic on a semi-Riemannian manifold with positive index is an infinite dimensional saddle point for the action integral.

The strong indefiniteness of the action integral makes it difficult to apply to semi-Riemannian manifolds the classical results of Critical Point Theory, based on the deformation by the gradient flow of the sublevels of a functional, which works very well on Riemannian manifolds. Indeed, critical points having Morse index equal to $+\infty$ do not change the homotopy type of the sublevels of a functional. (We are attaching infinite dimensional handles and the infinite dimensional unit sphere of a Hilbert space is contractible!)

A geodesic $z \in \Omega^{1,2}(p,q;\mathcal{M})$ is said to be nondegenerate if it is a nondegenerate critical point of the action integral f, i.e. if the second derivative defines an invertible linear operator on the tangent space $T_z\Omega^{1,2}(p,q;\mathcal{M})$. Since $f''(z)$ is a Fredholm operator of index 0, this is equivalent to requiring that the kernel of $f''(z)$ is trivial, and this is equivalent to saying that the Jacobi equation $D_s^2\zeta + R(\zeta,\dot{z})\dot{z} = 0$ has no nontrivial solution ζ such that $\zeta(0) = 0$, $\zeta(1) = 0$. Two points p and q of a semi-Riemannian manifold $(\mathcal{M},g)$ are said to be *nonconjugate* if any geodesic joining p and q is nondegenerate. From a variational point of view, the nonconjugacy of the points p and q means that the action integral (2) is a Morse function, i.e. all the critical points of f are nondegenerate. Using the Sard theorem it can be proved that all pairs of points in $\mathcal{M}$, except for a nowhere dense set, are nonconjugate, see [54].

We recall now the *Palais–Smale* compactness condition, which plays a basic role in infinite dimensional variational problems.

Definition 2. *Let $f : X \to \mathbf{R}$ be a C^1 functional defined on a Riemannian manifold (X,h) and let F be a closed subset of X, then the functional f satisfies the* Palais–Smale (PS) condition *on F if for any sequence $(x_m)_{m\in\mathbf{N}}$ of points of F, such that*

(i) $\{f(x_m)\}_{m\in\mathbf{N}}$ *is bounded*,
(ii) $\|\nabla f(x_m)\| \to 0$,

there exists a converging subsequence. Here $\|\cdot\|$ denotes the norm induced on the tangent bundle by the fixed Riemannian metric h on X.

For any $c \in \mathbf{R}$ we set

$$\begin{aligned} f^c &= \{x \in X \mid f(x) \leq c\}\,, \\ f_c &= \{x \in X \mid f(x) \geq c\}\ . \end{aligned} \tag{5}$$

Moreover, for any $a \leq b$ we set

$$f_a^b = \{x \in X \mid a \leq f(x) \leq b\}\ . \tag{6}$$

We present now the main results of Critical Point Theory. They are all based on two deformation theorems which show the relations between the change of homotopy type of the sublevels of a functional and the presence of critical points of the functional.

Theorem 1. *Let $f : (X, h) \to \mathbf{R}$ be a C^1 functional defined on a complete Riemannian manifold (X, h), let $a < b$ be two regular values of f and assume that there are no critical points in f_a^b and f satisfies the Palais–Smale condition on the closed set f_a^b.*

Then the sublevel f^a is a strong deformation retract of f^b, that is there exists a continuous homotopy $H : [0,1] \times f^b \to f^b$ such that

(i) $H(0, x) = x$, *for any* $x \in f^b$;
(ii) $H(t, y) = y$, *for any* $t \in [0,1]$ *and for any* $y \in f^a$;
(iii) $H(1, f^b) = f^a$.

The proof of this theorem can be found in [52, 64]. The idea is to construct the homotopy H using the flow lines of the gradient vector field ∇f of the functional f with respect to the Riemannian structure h of the manifold X. The absence of critical points of f on f_a^b and the Palais–Smale condition on the same set assure that the flow starting from f^b reaches the sublevel f^a in a finite time, remaining f^a fixed. This idea works only for C^2 functionals, for which the gradient is locally Lipschitz continuous and the Cauchy problem for the gradient flow has a unique solution. The proof for functionals of class C^1 is obtained using the notion of *pseudogradient* fields introduced by R. Palais.

The previous theorem claims that if there are no critical points in the strip f_a^b and the Palais–Smale condition holds, then f^b can be continuously deformed into f^a. On the other hand, if the Palais–Smale condition holds and the sublevels f^b and f^a are not homotopically equivalent, then a critical point exists in the strip f_a^b. The main abstract critical points theorems are based on topological arguments on the underlying manifold X and/or geometric assumptions on the function f, forcing two sublevels of the function to be homotopically different. Moreover, some assumption on X or f is necessary in order that f satisfies the Palais–Smale condition.

We point out that another situation can happen in this scenario: f satisfies (PS), the sublevel f^a is a deformation retract of f^b and in spite of this, there could be critical points of f in the strip f_a^b. This is a situation typical of

strongly indefinite functionals, that is functionals having critical points with Morse index equal to $+\infty$ as the action integral in semi-Riemannian manifolds with positive index. So existence results based on Theorem 1 do not well apply to semi-Riemannian geometry.

Theorem 1 can be extended, for the case that critical points of the functional are present, in the following way, see [52] for the proof.

Theorem 2. *Let $f : (X, h) \to \mathbf{R}$ be a C^1 functional defined on the complete Riemannian manifold (X, h), let $c \in \mathbf{R}$ and let $K_c = \{x \in X : f(x) = c, f'(x) = 0\}$ be the set of the critical points of f at the level c and assume that f satisfies* (PS).

Then, for any neighborhood U of K_c, there exists a positive number ϵ_0 such that for any $\epsilon \in]\,0, \epsilon_0\,[$ there exists a continuous homotopy $H_\epsilon : [0, 1] \times f^{c+\epsilon} \setminus U \to f^{c+\epsilon} \setminus U$ such that

(i) $H_\epsilon(0, x) = x$, *for any* $x \in f^{c+\epsilon} \setminus U$;
(ii) $H_\epsilon(t, y) = y$, *for any* $t \in [0, 1]$ *and* $y \in f^{c-\epsilon} \setminus U$;
(iii) $H_\epsilon(1, f^{c+\epsilon} \setminus U) = f^{c-\epsilon}$.

Theorems 1 and 2 are the basic tools for reducing results on the existence and the multiplicity of critical points of functionals bounded from below and satisfying the Palais–Smale condition. The following theorem is a consequence of Theorem 1.

Theorem 3. *Let $f : (X, h) \to \mathbf{R}$ be a C^1 functional defined on a complete Riemannian manifold (X, h), bounded from below and satisfying the Palais–Smale condition on X.*

Then the infimum is attained, i.e., there exists a point $x_0 \in X$ such that $f(x_0) = \inf_X f$.

If the topology of the manifold X is rich, we have a multiplicity result of critical points of a functional in terms of a topological invariant of the manifold, the *Ljusternik–Schnirelmann category* of X. For any topological space X, the Ljusternik–Schnirelmann category $\mathrm{cat}(X)$ is equal to the minimal number of closed and contractible subsets which cover X. If such a minimal number does not exist, it is $\mathrm{cat}(X) = +\infty$.

Theorem 4. *Let $f : (X, h) \to \mathbf{R}$ be a C^1 functional defined on a complete Riemannian manifold (X, h), bounded from below and satisfying the Palais–Smale condition.*

Then the functional f has at least $\mathrm{cat}(X)$ *critical points. Moreover, if* $\mathrm{cat}(X) = +\infty$, *then there exists a sequence x_n of critical points of f such that $f(x_n) \to +\infty$.*

The proof of this theorem was obtained by Ljusternik and Schnirelmann at the end of the twenties of the last century. If the manifold X is contractible, for instance if X is a Hilbert space, Theorem 4 reduces to the existence of a

minimum point for f. In order to obtain multiple critical points for f, some geometrical assumption on the functional f is needed.

Finally we present the results of Morse Theory for a functional f bounded from below and satisfying the Palais–Smale condition. Morse Theory gives more precise estimates for the critical points of a functional defined on a Hilbert manifold, in particular on the number of critical points having a fixed Morse index. However, in order to prove the results of Morse Theory, we have to pay two costs. Firstly we have to assume that the functional is of class C^2 and also all the critical points of f have to be nondegenerate, i.e. the functional f has to be a *Morse function*. We first state a result which is a refinement of Theorems 1 and 2. It gives more precise information on the topological change of the sublevels in the presence of a nondegenerate critical point, see [56].

Theorem 5. *Let $f : (X, h) \to \mathbf{R}$ be a C^2 functional defined on a complete Riemannian manifold (X, h), let $a < b$ be two regular values of f and assume that on the closed strip f_a^b there is only a nondegenerate critical point x with $c = f(x) \in]a, b[$.*

Then the sublevel f^b is homotopically equivalent to the topological space obtained by the connected sum of the sublevel f^a with a k-dimensional handle B^k attached to f^a at the boundary S^{k-1}, where $k = m(x, f)$ is the Morse index of the critical point x, B^k is the k-dimensional unit disk and S^{k-1} is its boundary, the $(k-1)$-dimensional unit sphere. In particular, if $k = +\infty$ (as in semi-Riemannian geometry), f^a is a deformation retract of f^b.

The last statement comes from the fact that, unlike in finite dimension, the infinite dimensional unit ball B^∞ is contractible onto its boundary S^∞, so that the topological pair (B^∞, S^∞) is trivial. In particular, nondegenerate critical points with Morse index equal to $+\infty$, for instance the geodesics of semi-Riemannian manifolds with positive index, do not affect the topological properties of sublevels of the functionals and their existence cannot be deduced by change-of-homotopy arguments. In a certain sense, critical points of strongly indefinite functionals are invisible to continuous homotopies.

Let (A, B) be a topological pair, that is A is a topological space and B is a subspace of A, and let $\mathcal{K}$ be a field. For any $k \in \mathbf{N}$, $H_k(A, B; \mathcal{K})$ denotes the k-th relative homology group (with coefficients in $\mathcal{K}$) of the pair (A, B) (cf. [63]). Since $\mathcal{K}$ is a field, the homology group $H_k(A, B; \mathcal{K})$ is a vector space and its dimension $\beta_k(A, B; \mathcal{K}) \in \mathbf{N} \cup \{+\infty\}$ is called the k-th *Betti number* of (A, B) (with respect to $\mathcal{K}$). The Poincaré polynomial of the pair (A, B) is defined by setting

$$\mathcal{P}(A, B; \mathcal{K})(r) = \sum_{k=0}^{\infty} \beta_k(A, B; \mathcal{K})\, r^k \ .$$

In general $\mathcal{P}$ is a formal series whose coefficients are positive cardinal numbers belonging to $\mathbf{N} \cup \{+\infty\}$.

We state now the Morse relations, the Morse inequalities and the total Betti number formula for a Morse functional, bounded from below and such that the Morse index of any critical point is finite. They relate the numbers of critical points of the functionals to the Betti numbers of the manifold. For the proof see [11, 52].

Theorem 6. *Let $f : X \to \mathbf{R}$ be a C^2 functional defined on a complete Riemannian manifold (X, h). Assume that f is bounded from below and satisfies the Palais–Smale condition on X. Moreover, assume that all critical points of f are nondegenerate and that the Morse index $m(x, f)$ of any critical point x of f is finite.*

Then for any field $\mathcal{K}$ there exists a formal series $Q(r)$, whose coefficients are positive cardinal numbers, such that

$$\sum_{x \in K(f)} r^{m(x,f)} = \mathcal{P}(X, \mathcal{K})(r) + (1 + r)\, Q(r) \; . \tag{7}$$

Moreover, let for any $k \in \mathbf{N}$, $\beta_k(X; \mathcal{K})$ be the k-th Betti number of the manifold X with respect to the field $\mathcal{K}$ and denote by $M(f, k)$ the number of critical points x of f such that $m(x, f) = k$. Then

$$M(f, k) \geq \beta_k(X; \mathcal{K}) \; . \tag{8}$$

Finally, let $K(f)$ be the set of the critical points of f, and denote by $\mathcal{B}(X, \mathcal{F})$ the total Betti number of f with respect to the field $\mathcal{K}$ defined as

$$\mathcal{B}(X, \mathcal{K}) = \sum_{k=0}^{\infty} \beta_k(X, \mathcal{K}) \; .$$

Then the number $\sharp K(f)$ of the critical points of f satisfies the relation

$$\sharp K(f) = \mathcal{B}(X, \mathcal{K}) + 2\, Q(1) \; . \tag{9}$$

Notice that, under the assumptions of the previous theorem, the number of critical points of the functional f is countable, because nondegenerate critical points are isolated, and the Palais–Smale condition holds on the whole manifold X, so that sums are well defined.

The abstract critical point theorems stated above hold for functionals bounded from below, and they well apply to the action integral in Riemannian geometry. On the other hand, the action integral in semi-Riemannian geometry is unbounded both from below and from above. The existence and the multiplicity of critical points of unbounded functionals have been considered in Nonlinear Analysis, trying to find variational solutions of semilinear elliptic or hyperbolic partial differential equations, Hamiltonian systems and geodesics on Lorentzian geometry. The classical results on this topic are the *Mountain Pass Theorem*, the *Saddle Point Theorem* and the socalled *Linking*

Theorems which unify this class of results. We refer to [52, 64] for statements and proofs. Moreover, a generalization of the Ljusternik–Schnirelmann category, the socalled *relative category*, has been introduced by some authors in various (but essentially equivalent) ways in order to obtain multiplicity results for strongly indefinite functionals. We refer to [23, 28] for the relative category and to [34, 13] for applications to geodesics on Lorentzian manifolds. Here we state a simplified version of the *Saddle Point Theorem* due to P.H. Rabinowitz.

Theorem 7. *Let (X, h) be an infinite dimensional complete Riemannian manifold, let H be a Hilbert space, let F_0 be a finite dimensional subspace of H and let $F = e_0 + H_0$ be a finite dimensional affine submanifold of H, with $e_0 \in H$. Finally, set $\Omega = X \times F$. Let $f\colon \Omega = X \times F \to \mathbf{R}$ be a C^1 functional such that:*

- *the functional f satisfies the Palais–Smale condition;*
- *there exists a point $z^* = (x^*, t^*) \in \Omega$ such that $f(x^*, t) \to -\infty$ as $\|t\| \to +\infty$ and $f(x, t^*) \to +\infty$ as $\|x\| \to +\infty$.*

Then the functional f admits a critical point which is a saddle point.

The functional f in Theorem 7 is *not* strongly indefinite and the found critical point is a finite dimensional saddle point having a finite Morse index. For this reason Theorem 7 cannot be applied directly to the action integral of a semi-Riemannian manifold, but only modulo finite dimensional approximations.

3 Geodesics on Riemannian Manifolds

The results stated in the previous section can be applied to the action integral of a complete Riemannian manifold. We obtain existence and multiplicity results and a Morse Theory for geodesics on a Riemannian manifold. These results were already the core of the results of Morse, Ljusternik and Schnirelmann, obtained essentially using finite dimensional reductions of the problem (see the classical book of J. Milnor [54] on Morse Theory). The infinite dimensional approach using Hilbert manifolds, gradient flows and the Palais–Smale condition was introduced by R. Palais in the celebrated paper [56], see also [11, 48].

Let $(\mathcal{M}, g)$ be a *complete Riemannian manifold*, let p and q be two points of $\mathcal{M}$ and consider the action integral $f(x) = \int_0^1 g(x(s))[\dot{x}(s), \dot{x}(s)]ds$ on the manifold $\Omega^{1,2}(p, q; \mathcal{M})$. The functional f is bounded from below. Moreover, the completeness of $(\mathcal{M}, g)$ allows to prove that f satisfies the (PS) condition, see [56]. By Theorem 3, there exists a minimum of f, so there exists a minimal geodesic joining p and q. We have obtained a variational proof of the geodesic connectedness of a complete Riemannian manifold, which is usually proved as a consequence of the well-known Hopf–Rinow Theorem

in Riemannian geometry. The variational methods work very well to prove a multiplicity result. Indeed, E. Fadell and S. Husseini have proved in [22] that the Ljusternik–Schnirelmann category $\mathrm{cat}(\Omega^{1,2}(p,q;\mathcal{M}))$ is equal to $+\infty$ whenever the manifold $\mathcal{M}$ is not contractible to a point. Then, if $\mathcal{M}$ is non-contractible, for any complete Riemannian metric g and for any pair of points p and q on $\mathcal{M}$, there exist infinitely many geodesics joining p and q and there exists a sequence (x_n) of such geodesics such that the action integral $f(x_n)$ tends to $+\infty$. This result was already proved by Serre [62] in the case of a compact and simply connected manifold, using spectral sequences to study the topology of $\Omega^{1,2}(p,q;\mathcal{M})$.

If the Riemannian manifold is complete and the p and q are nonconjugate (a condition which holds almost surely), the action integral satisfies all the assumptions of the abstract Theorem 6. Indeed, f is bounded from below, it satisfies the Palais–Smale condition, and by the nonconjugacy of p and q all the critical points of f are nondegenerate. Finally, since a Riemannian metric is positive definite, by (4) it follows that the Morse index of any geodesic is finite. So the Morse relations (7), the Morse inequalities (8) and the formula (9) for the total number of geodesics hold for the geodesics joining p and q.

The variational properties of the action integral of a complete Riemannian manifold are completely described. Moreover, since the infinite dimensional manifold $\Omega^{1,2}(p,q;\mathcal{M})$ is homotopically equivalent to the based loop space $\Omega(\mathcal{M})$, their homology groups are isomorphic, so we have a full relation between the differential structure of the geodesics of the complete Riemannian metric g and the topological structure of the manifold $\mathcal{M}$.

Assume now that the manifold $\mathcal{M}$ is contractible, so the infinite dimensional spaces $\Omega(\mathcal{M})$ and $\Omega^{1,2}(p,q;\mathcal{M})$ are contractible and their category is equal to 1. Then, for any field $\mathcal{K}$, the only nonzero Betti number of $\Omega(\mathcal{M})$ is $\beta_0(\Omega(\mathcal{M});\mathcal{K}) = 1$, so the total Betti number $\mathcal{B}(\mathcal{M};\mathcal{K})$ of the based loop space $\Omega(\mathcal{M})$ is equal to 1 . Then both Ljusternik–Schnirelmann Theory and Morse Theory give the existence of at least one geodesic joining two points p and q on $\mathcal{M}$, the minimal one. Moreover the total number $\sharp\mathcal{G}(p,q)$ of geodesics joining p and q satisfies the relation

$$\sharp\mathcal{G}(p,q) = \mathcal{B}(\mathcal{M},\mathcal{K}) + 2\,Q(1) = 1 + 2\,Q(1)\ .$$

So we have shown that if $(\mathcal{M},g)$ is complete, the manifold $\mathcal{M}$ is contractible and p and q are nonconjugate, then the number of geodesics joining p and q is infinite (when $Q(1) = +\infty$) or it is *odd* (when $Q(1) < +\infty$). An example of a manifold with three geodesics joining two nonconjugate points is the revolution paraboloid $z = x^2 + y^2$. Examples in which the number of geodesics joining two points is greater than one are interesting in the study of the geometric causes of the so called *multiple image effect*, studied in Astrophysics to describe the *gravitational lens effect*. In particular, Morse Theory for Riemannian metrics gives a proof of the oddity of the number of images in gravitational lensing in a *conformally static spacetime*. We refer to the papers [36, 37, 38, 40] for extensions in these direction.

4 Geodesics on Stationary Lorentzian Manifolds

We consider now a semi-Riemannian manifold $(\mathcal{M}, g)$ and the action integral $\int_0^1 g(z(s))[\dot{z}(s), \dot{z}(s)]ds$ on the infinite dimensional manifold $\Omega^{1,2}(p, q; \mathcal{M})$, whose critical points are the geodesics joining p and q. We cannot apply the abstract critical point theorems for functionals bounded from below because the functional $f(z)$ is unbounded. Moreover, the functional f does not satisfy, in general, the Palais–Smale condition and the Morse index of any geodesic is equal to $+\infty$.

However there is a class of Lorentzian manifolds where the methods of Riemannian manifolds work. This is the class of *stationary Lorentzian manifolds*. We recall some definitions. Let $(\mathcal{M}, g)$ be a semi-Riemannian manifold. A C^1 vector field $Y(z)$ on $\mathcal{M}$ is called a *Killing vector field* for g if the Lie derivative of the metric g with respect to Y vanishes, or equivalently, if for any pair of vector fields W_1 and W_2 on $\mathcal{M}$, it is

$$g(z)[\nabla_{W_1} Y, W_2] + g(z)[\nabla_{W_2} Y, W_1] = 0 \,,$$

where ∇ denotes the Levi–Civita connection for the metric g. This definition is equivalent to claiming the strong property that all stages of the flow of Y are isometries for $(\mathcal{M}, g)$, see [55, 68] for details. The main property of a Killing field with respect to geodesics is the following: Let Y and $z(s)$ be respectively a Killing field and a geodesic for $(\mathcal{M}, g)$, then a conservation law holds for $z(s)$, because

$$\frac{\mathrm{d}}{\mathrm{d}s} g(z(s))[\dot{z}(s), Y(z(s))] = 0 \,. \tag{10}$$

Definition 3. *A Lorentzian manifold is called* stationary *if it admits a smooth vector field $Y(z)$ which is both Killing and timelike, that is $g(z)[Y(z), Y(z)] < 0$, for any $z \in \mathcal{M}$. A stationary Lorentzian manifold is called* static *if the orthogonal distribution $Y^\perp$ to the timelike Killing field Y is integrable* (*cf.* [55]).

A wide class of stationary and static Lorentzian manifolds are the *standard* ones. A Lorentzian manifold $(\mathcal{M}, g)$ is standard stationary if the manifold $\mathcal{M}$ is diffeomorphic to a product manifold $\mathcal{M}_0 \times \mathbf{R}$ and, setting $z = (x, t) \in \mathcal{M}$ with $x \in \mathcal{M}_0$ and $t \in \mathbf{R}$, the metric g in the coordinates (x, t) takes the following form: for any $z = (x, t) \in \mathcal{M}$ and for any $\zeta = (\xi, \tau) \in T_z\mathcal{M} = T_x\mathcal{M}_0 \times \mathbf{R}$,

$$g(z)[\zeta, \zeta] = \langle \xi, \xi \rangle + 2\langle \delta(x), \xi \rangle \tau - \beta\,(x)\, \tau^2 \,, \tag{11}$$

where $\langle \cdot, \cdot \rangle$ is a Riemannian metric on $\mathcal{M}_0$, $\delta(x)$ is a smooth vector field on $\mathcal{M}_0$ and $\beta\,(x)$ is a smooth positive scalar field on $\mathcal{M}$. Moreover, $(\mathcal{M}, g)$ is a *standard static Lorentzian manifold* if and only if the vector field $\delta(x)$ in (11) is zero. Notice that ∂_t is a timelike Killing vector field with respect to g.

Many physically relevant spacetimes in Relativity are stationary or static. The *Minkowski spacetime* of Special Relativity, the *Schwarzschild spacetime* outside the event horizon and the *Reissner–Nordström spacetime* outside the first event horizon are examples of standard static spacetimes, while the *Kerr spacetime* outside the stationary limit surface is an example of standard stationary spacetime, see [46]. On the other hand, there are also interesting examples of nonstandard stationary metrics. The *Gödel spacetime* is stationary but nonstandard, because it cannot be written in the form of (11) with a positive definite metric $\langle \cdot, \cdot \rangle$. Moreover, there exists a stationary metric on the 3-sphere S^3 and static metrics on manifolds topologically equivalent to a torus.

We fix now a stationary Lorentzian manifold $(\mathcal{M}, g)$ with a timelike smooth Killing vector field $Y(z)$ and we fix two points p and q on $\mathcal{M}$. The action integral $\int_0^1 g(z(s))[\dot{z}(s), \dot{z}(s)]ds$ is unbounded on the infinite dimensional manifold $\Omega^{1,2}(p, q; \mathcal{M})$. However, a new variational principle for geodesics joining p and q can be proved. Indeed, the property of $Y(z)$ to be a Killing vector field gives the existence of a *natural constraint* for the functional f: there is a submanifold $\mathcal{N}(p, q)$ of $\Omega^{1,2}(p, q; \mathcal{M})$ such that the restriction of f to $\mathcal{N}(p, q)$ has the same critical points as the free functional f on $\Omega^{1,2}(p, q; \mathcal{M})$, that is the geodesics joining p and q. In other words, the Killing property of $Y(z)$ allows to kill (!) the time directions on $\Omega^{1,2}(p, q; \mathcal{M})$ which are responsible for the strongly indefinite nature of the action integral on a Lorentzian manifold. The manifold $\mathcal{N}(p, q)$ is defined as follows:

Definition 4. *Let $(\mathcal{M}, g, Y)$ be a stationary spacetime, p and q two points on $\mathcal{M}$. The* natural constraint *of the action integral on $\Omega^{1,2}(p, q; \mathcal{M})$ is defined as*

$$\mathcal{N}(p, q) = \{z \in \Omega^{1,2}(p, q; \mathcal{M}) \ : \\ g(z)[\dot{z}, Y(z)] \text{ is constant almost everywhere on } [0, 1]\} \ .$$

Notice that by (10) any geodesic joining p and q belongs to the natural constraint $\mathcal{N}(p, q)$. Moreover, again the Killing property of $Y(z)$ allows to prove the following variational principle.

Theorem 8. *Let $(\mathcal{M}, g, Y)$ be a stationary spacetime, p and q two points on $\mathcal{M}$ and assume that the natural constraint $\mathcal{N}(p, q)$ is not equal to the empty set. Then the set $\mathcal{N}(p, q)$ is an infinite dimensional* smooth *submanifold of $\Omega^{1,2}(p, q; \mathcal{M})$. Let $J(z)$ be the restriction of the action integral to $\mathcal{N}(p, q)$, then a curve $z \in \mathcal{N}(p, q)$ is a critical point of J if and only if z is a geodesic joining p and q, that is z is a critical point of the free functional f on $\Omega^{1,2}(p, q; \mathcal{M})$.*

This variational principle was proved for an arbitrary stationary Lorentzian manifold in [43]. It was first proved in [5] and [32] respectively for standard static and standard stationary Lorentzian manifolds. In the standard case $\mathcal{N}(p, q)$ is smooth since it is diffeomorphic to the infinite dimensional manifold $\Omega^{1,2}(p_0, q_0; \mathcal{M}_0)$, where p_0 and q_0 are projections of p and q to $\mathcal{M}_0$, see also

(11). We point out that sometimes the natural constraint $\mathcal{N}(p,q)$ is equal to the empty set, see [43] for an example. In the standard case, $\mathcal{N}(p,q)$ is not empty if $\mathcal{M}$ is connected. The variational principle stated above does not hold if Y is only a timelike vector field; it strongly relies on the Killing nature of Y.

We state now the main theorem on the geodesic connectedness of a stationary Lorentzian manifold. We first give a definition, see [43].

Definition 5. *Let p and q be two points of $\mathcal{M}$ and $c \in \mathbf{R}$. We say that the action integral f is* c-precompact *on $\mathcal{N}(p,q)$ if any sequence $(z_n)_{n\in\mathbf{N}}$ of curves in $\mathcal{N}(p,q)$ such that*

$$f(z_n) \leq c\,, \qquad \forall n \in \mathbf{N}$$

contains a subsequence which converges uniformly to a curve z in $\mathcal{N}(p,q)$. We say that f is precompact *on $\mathcal{N}(p,q)$ if it is c-precompact on $\mathcal{N}(p,q)$ for any real number c.*

The definition of c-precompactness is a generalization of the notion of completeness of a Riemannian manifold, even if it requires a property of the functional f on an infinite dimensional manifold, rather than a property of the physical manifold $(\mathcal{M},g)$. The c-precompactness is satisfied by the action integral on the manifold of curves joining two points in a complete Riemannian manifold. We state now the main theorem on the geodesic connectedness of a stationary Lorentzian manifold, see [43].

Theorem 9. *Let $(\mathcal{M},g,Y)$ be a stationary Lorentzian manifold and let p and q two points of $\mathcal{M}$. Assume that $\mathcal{N}(p,q)$ is nonempty and the action integral $f(z)$ is precompact on $\mathcal{N}(p,q)$. Let $J(z)$ be the functional obtained by the restriction of $f(z)$ to $\mathcal{N}(p,q)$.*

Then the functional J is bounded from below and satisfies the Palais–Smale condition. So J has a minimum which is a geodesic joining p and q. If the assumptions above hold for any p and q, then $(\mathcal{M},g,Y)$ is geodesically connected.

Moreover, if f is precompact on $\mathcal{N}(p,q)$, the Killing field is complete and the manifold $\mathcal{M}$ is not contractible to a point, then infinitely many geodesics joining p and q exist and a sequence $(z_n)_{n\in\mathbf{N}}$ of such geodesics satisfies $f(z_n) \to +\infty$ as $n \to \infty$.

This theorem is very general in its statement. In the case of standard stationary manifolds, suitable assumptions on the components $\langle\cdot,\cdot\rangle$, $\delta(x)$ and $\beta(x)$ allow to obtain the geodesic connectedness and the existence of infinitely many geodesics, see [5, 32].

Theorem 10. *Assume that the stationary Lorentzian manifold $(\mathcal{M},g)$ is of standard type and assume that the Riemannian metric $\langle\cdot,\cdot\rangle$ on $\mathcal{M}_0$ is complete, the vector field $\delta(x)$ satisfies*

$$\sup\{\langle\delta(x),\delta(x)\rangle, x \in \mathcal{M}_0\} < +\infty$$

and the scalar field $\beta(x)$ *satisfies*

$$0 < L \leq \beta(x) \leq M < +\infty$$

for some positive constants L, M *and for any* $x \in \mathcal{M}_0$. *Then for any pair of points* p *and* q *of* $\mathcal{M}$, *the functional* f *is precompact on* $\mathcal{N}(p,q)$ *and the standard manifold* $(\mathcal{M}, g)$ *is geodesically connected. Moreover, if the manifold* $\mathcal{M}$ *is noncontractible, then infinitely many geodesics joining* p *and* q *exist and a sequence* $(z_n)_{n\in\mathbf{N}}$ *of such geodesics satisfies* $f(z_n) \to +\infty$. *Finally, if the metric is standard static, all the results above hold assuming for the scalar field* $\beta(x)$ *only that it is bounded from above.*

The assumptions on δ and β can be weakened, assuming that δ has sublinear growth and β has subquadratic growth at infinity. Moreover, the geodesic connectedness can be proved also in the standard static case when the scalar field has quadratic growth at infinity, see for instance [26]. For superquadratic growths there are counter-examples to the geodesic connectedness as the anti-de Sitter spacetime, see [57].

For nonstandard static manifolds, the following result is proved in [17].

Theorem 11. *Any compact static Lorentzian manifold* $(\mathcal{M}, g, Y)$ *is geodesically connected. Moreover, in any homotopy class of curves joining two points of the manifold, there is a geodesic minimizing the action integral on the intersection of the homotopy class with the natural constraint.*

As Miguel Sánchez has pointed out, the geodesic connectedness can be realized for any pair of points by a timelike geodesic; that is, a compact static manifold is *totally vicious*.

The assumption on the static manifold can be weakened. The geodesic connectedness of a static Lorentzian manifold holds also if the manifold $\mathcal{M}$ is not necessarily compact, but the function $\beta(z) = -g(z)[Y(z), Y(z)]$ has at most quadratic growth at infinity, and the following Riemannian metric g_R on $\mathcal{M}$ associated with g and Y is complete:

$$g_R(z)[v,w] = g(z)[v,w] + 2\,\frac{g(z)[Y(z),v]g(z)[Y(z),w]}{\beta(z)}\,.$$

It is not known if the results stated above for static metrics hold for stationary manifold. For instance, the geodesic connectedness of a compact stationary manifold is an open problem. A generalization to *semi-Riemannian manifolds* of index k admitting k timelike, linearly independent Killing vector fields satisfying some other technical assumptions, is contained in [44]. In the paper [14] it is studied the geodesic connectedness of the *spacetimes of Gödel type*. It is a class of stationary spacetimes, including the well-known solution of the Einstein equation obtained by Kurt Gödel, that contain closed

timelike curves. In particular it is shown in [44] that the Gödel spacetime is geodesically connected. Finally in the paper [3] the geodesic connectedness of standard static spacetimes is studied under some geometric assumptions.

We consider now Morse Theory for geodesics on stationary spacetimes. The reduction to the natural constraint permits to develop a Morse Theory for the geodesics joining two nonconjugate points on a stationary Lorentzian manifold. We first state the following result which shows that the restriction of the action integral to the natural constraint is *not* strongly indefinite.

Theorem 12. *Let $(\mathcal{M}, g, Y)$ be a stationary spacetime and let $z\colon [0,1] \to \mathcal{M}$ be a geodesic for $\mathcal{M}$, with $z(0) = p$ and $z(1) = q$. Let $\mathcal{N}(p,q)$ be the natural constraint for the action integral f on $\Omega^{1,2}(p,q;\mathcal{M})$ and let J be the restriction of f to $\mathcal{N}(p,q)$. Finally consider the continuous linear operator $J''(z)\colon T_z\mathcal{N}(p,q) \to T_z\mathcal{N}(p,q)$ induced on $T_z\mathcal{N}(p,q)$ by the Hessian $H_J(z)$ of the critical point z of J.*

Then $J''(z)$ is a Fredholm operator of index 0 and it is a compact perturbation of a positive definite isomorphism on $T_z\mathcal{N}(p,q)$. In particular the Morse index $m(z,J)$ and the augmented Morse index $m^*(z,J)$ are finite. Moreover, z is a nondegenerate critical point of J if and only if q is nonconjugate to p along z. In this case it is $m(z,J) = m^*(z,J)$.

Collecting all the results above, we can prove the Morse relations for the geodesics joining two nonconjugate points on a stationary Lorentzian manifold. We recall that, on a semi-Riemannian manifold, the set of pairs of nonconjugate points has measure equal to 0, so the next theorem holds for almost all pairs of points, see [42] for the proof in the general case of stationary metrics, while for a simpler one in the standard cases see [9, 35].

Theorem 13. *Let $(\mathcal{M}, g, Y)$ be a stationary spacetime and let p and q be two points of the manifold. Let $\mathcal{N}(p,q)$ be the natural constraint for the action integral on $\Omega^{1,2}(p,q;\mathcal{M})$ and let J be the restriction of f to $\mathcal{N}(p,q)$. Then the functional J is a Morse functional if and only if p and q are nonconjugate.*

Assume now that p and q are nonconjugate. Moreover, assume that the timelike Killing field Y is complete and f is precompact on $\mathcal{N}(p,q)$. Then, for any field $\mathcal{K}$ there exists a formal series $Q(r)$, whose coefficients are positive cardinals, such that, denoting by $\mathcal{G}(p,q)$ the set of the geodesics joining p and q, we have

$$\sum_{z \in \mathcal{G}(p,q)} r^{m(z,J)} = \mathcal{P}(\Omega(\mathcal{M}), \mathcal{K}) + (1+r)\, Q(r)\ ,$$

where $\Omega(\mathcal{M})$ is the based loop space of the manifold $\mathcal{M}$. Moreover the Morse inequalities (8) and the total Betti number formula (9) hold.

5 Geodesics on Splitting Lorentzian Manifolds

In this section we present some results on the geodesics joining two points for the class of orthogonally splitting Lorentzian manifolds.

Definition 6. *A Lorentzian manifold* $(\mathcal{M}, g)$ *is said to be* orthogonally splitting *if* $\mathcal{M} = \mathcal{M}_0 \times \mathbf{R}$*, where* $\mathcal{M}_0$ *is a smooth connected manifold, and the metric* g *has the following form. For any* $z = (x,t) \in \mathcal{M}$ *and for any* $\zeta = (\xi, \tau) \in T_z\mathcal{M} = T_x\mathcal{M}_0 \times \mathbf{R}$*,*

$$g(z)[\zeta,\zeta] = \langle \alpha(x,t)\xi, \xi\rangle - \beta(z)\,\tau^2\,, \tag{12}$$

where $\langle\cdot,\cdot\rangle$ *is a Riemannian metric on* $\mathcal{M}_0$*,* $\alpha(x,t)$ *is a positive linear operator on* $T_x\mathcal{M}_0$*, smoothly depending on* z*, and* $\beta(z)$ *is a smooth positive scalar field on* $\mathcal{M}$*.*

The notion of orthogonally splitting manifolds can be introduced in a more intrinsic way. A Lorentzian manifold $(\mathcal{M}, g)$ is orthogonally splitting if it is *stably causal* (see [46]) and has a smooth time function $T\colon \mathcal{M} \to \mathbf{R}$ such that the smooth timelike vector field ∇T is complete. Moreover, we have to assume that all the level hypersurfaces $T^{-1}(c)$ are diffeomorphic, so that the manifold $\mathcal{M}$ is foliated into a cross product $\mathcal{M} = \mathcal{M}_0 \times \mathbf{R}$, where $\mathcal{M}_0 = T^{-1}(0)$.

Many physically relevant classes of spacetimes are orthogonally splitting, as the Robertson–Walker and the generalized Robertson–Walker spacetimes, see [55, 24]. A classical result of Geroch [31] states that any *globally hyperbolic* Lorentzian manifold is homeomorphic to an orthogonally splitting one. Very recently it has been shown by Bernal and Sánchez [10] that the homeomorphism can be replaced by a diffeomorphism.

We fix now an orthogonally splitting Lorentzian manifold $(\mathcal{M}, g)$ and two points $p = (x_0, t_0)$ and $q = (x_1, t_1)$ on $\mathcal{M}$; our aim is to study the geodesics for g joining p and q. The action integral is expressed now by

$$f(z) = \int_0^1 g(z)[\dot z, \dot z]\,ds = \int_0^1 \langle \alpha(x,t)\dot x, \dot x\rangle\,ds - \int_0^1 \beta(z)\dot t^2\,ds\,, \tag{13}$$

defined on the manifold $\Omega^{1,2}(p,q;\mathcal{M}) \equiv \Omega^{1,2}(x_0,x_1;\mathcal{M}_0) \times \Omega^{1,2}(t_0,t_1;\mathbf{R})$.

The action functional f is unbounded both from below and from above as it clearly appears from (13). On the other hand, under simple assumptions on the coefficients $\alpha(z)$ and $\beta(z)$ (for instance, if they are bounded below away from 0 and bounded above away from $+\infty$), the functional f has the *saddle point geometry*. Unfortunately Theorem 7 cannot be directly applied to f for two reasons: first of all, the action functional f does not satisfy the Palais–Smale condition. Moreover, the saddle geometry of f is not finite dimensional as in Theorem 7, because the manifold $\Omega^{1,2}(t_0,t_1;\mathbf{R})$ is infinite dimensional. Indeed f is a strongly indefinite functional. In spite of this the following theorem has been proved in [7].

Theorem 14. *Let us assume that the orthogonally splitting Lorentzian manifold $(\mathcal{M}, g)$ satisfies the following properties:*

A_1) *The Riemannian manifold $(\mathcal{M}_0, \langle\cdot,\cdot\rangle)$ is complete;*

A_2) *there exists $\lambda > 0$ such that, for any $z = (x,t) \in \mathcal{M}$, and for any $\xi \in T_x\mathcal{M}_0$,*

$$\langle \alpha(z)\xi, \xi\rangle \geq \lambda \langle \xi, \xi\rangle \ ;$$

A_3) *there exists two positive constants $0 < \nu \leq M$ such that, for any $z \in \mathcal{M}$,*

$$\nu \leq \beta\,(z) \leq M \ ;$$

A_4) *there exists $L > 0$ such that, for any $z \in \mathcal{M}$,*

$$|\langle \alpha_t(z)\xi, \xi\rangle| \ \leq \ L\,\langle \xi, \xi\rangle\,, \qquad |\beta_t\,(z)\,| \ \leq \ L\,,$$

where α_t and β_t denote respectively the partial derivative, with respect to t, of α and β;

A_5)

$$\text{limsup}_{t\to+\infty}\langle \alpha_t(x,t)\xi, \xi\rangle \leq 0\,,$$

$$\text{liminf}_{t\to-\infty}\langle \alpha_t(x,t)\xi, \xi\rangle \geq 0\,,$$

uniformly in $x \in \mathcal{M}_0$ and $\xi \in T_x\mathcal{M}_0$, $\langle \xi, \xi\rangle = 1$.

Then the Lorentzian manifold $(\mathcal{M}, g)$ is geodesically connected.

The proof of Theorem 14 is of a variational nature. The Rabinowitz Saddle Point Theorem is applied to a family of functionals which approximate the action integral f, satisfy the Palais–Smale condition and have the geometry of a finite dimensional saddle point. A priori estimates on the critical points of these functionals allow to pass to the limit and to find a critical point of f. Assumptions A_1)–A_4) are needed to prove the Palais–Smale condition, while assumption A_5) is essential to prove the a priori estimates and in the limit process of the critical points of the approximating functionals. The assumptions A_1)–A_5) have been weakened in the paper [50], where assumption A_5) is replaced by the existence of a family of covering subsets of $\mathcal{M}$ which are invariant by the gradient flow of the time function T and whose boundaries satisfy a convexity assumption. In a certain sense assumption A_5) can be interpreted as a convexity at infinity of the metric g.

The results proved above for orthogonally splitting Lorentzian manifolds can be easily extended to orthogonally splitting semi-Riemannian manifolds of index k. In this case the manifold is a product $\mathcal{M}_0 \times \mathbf{R}^k$ and the metric is positive definite on the tangent bundle of $\mathcal{M}_0$ and negative definite on the tangent bundle of $\mathbf{R}^k$. Some results on the geodesic connectedness on generalized Robertson–Walker spacetimes, not using variational methods but integrating the geodesic equations, have been obtained in [24].

If the topology of the manifold $\mathcal{M}$ is nontrivial, we have the following multiplicity result, see [34, 13] based on abstract multiple critical point theorems for unbounded functionals obtained using the relative category and a theorem of Fadell and Husseini [23] on the relative category of the based loop space of a noncontractible manifold.

Theorem 15. *Let $(\mathcal{M}, g)$ be an orthogonally splitting Lorentzian manifold satisfying A_1)–A_5) and assume that the manifold $\mathcal{M}$ is not contractible to a point.*

Then, for any pair of points $p = (x_0, t_0)$ and $q = (x_1, t_1) \in \mathcal{M} = \mathcal{M}_0 \times \mathbf{R}$, there exist infinitely many geodesics joining them. Moreover there exists a sequence (z_m) of such geodesics such that $f(z_m) \to +\infty$.

The development of a Morse Theory for geodesics on a splitting Lorentzian manifold is a delicate problem. In the case of stationary metrics, Morse Theory was obtained owing to the introduction of the natural constraint where the action integral is not strongly indefinite and the Morse index of any geodesic is finite. But now we have no natural constraints. In a certain sense the geodesic problem in a nonstationary spacetime is genuinely strongly indefinite.

Morse Theory for strongly indefinite functionals has been the object of several studies in the last years and many applications to Hamiltonian systems, wave maps and symplectic geometry have been obtained. In particular, the definition of a new index for a critical point of a functional such that the second differential at the critical point is a Fredholm operator of index 0 has been studied by many authors [1, 2, 19, 20].

We define here a *relative index* and an *augmented relative index* for a class of bilinear forms on a Hilbert space. Let H be a real Hilbert space and let $a\colon H \times H \longrightarrow \mathbf{R}$ be a continuous, symmetric, bilinear form on H such that $a = a_0 + k$, where a_0 is a continuous, symmetric, nondegenerate bilinear form on H and k is a bilinear compact form. Let A and A_0 and K be the linear operators on H induced by the forms a, a_0 and k. Notice that $A = A_0 + K$, the operator A_0 is an isomorphism and K is a compact operator, so the operator A is a Fredholm operator of index 0. We denote by ker(a) the kernel of the bilinear form a which is equal to the kernel of the operator associated with A. Notice that ker(a) is a finite dimensional subspace of H. Moreover we denote by $V^+(A)$ and $V^-(A)$ the maximal A-invariant subspaces on which A is respectively positive definite and negative definite. Analogously, the A_0-invariant subspaces $V^+(A_0)$ and $V^-(A_0)$ on which A_0 is positive definite and negative definite are defined. The index $j(a, a_0)$ of the bilinear form a relative to a_0 is defined by setting:

$$j(a, a_0) = \dim(\mathrm{V}^-(\mathrm{A}) \cap \mathrm{V}^+(\mathrm{A}_0)) - \dim(\mathrm{V}^+(\mathrm{A}) \cap (\mathrm{V}^-(\mathrm{A}_0))) \,. \tag{14}$$

Moreover, the augmented index $j^*(a, a_0)$ of a with respect to a_0 is

$$j^*(a, a_0) = j(a, a_0) + \dim\big(\ker(\mathrm{a})\big) \,. \tag{15}$$

The indices $j(a, a_0)$ and $j^*(a, a_0)$ are relative integers. They coincide if and only if also the bilinear form a is nondegenerate. Moreover, if a_0 is a positive definite bilinear form, the relative index $j(a, a_0)$ is equal to the Morse index of a, the maximal dimension of a subspace where a is negative definite.

The relative index introduced is well defined for geodesics on a semi-Riemannian manifold. Let $(\mathcal{M}, g)$ be a semi-Riemannian manifold and let $z(s)\colon [0,1] \to \mathbf{R}$ be a geodesic for g with $z(0) = p$ and $z(1) = q$, so z is a critical point of the action integral $f(z) = \int_0^1 g(z(s))[\dot{z}(s), \dot{z}(s)]ds$ on the manifold $\Omega^{1,2}(p, q; \mathcal{M})$. The Hessian of f at z is the bilinear form on $T_z\Omega^{1,2}(p, q; \mathcal{M})$ given by

$$H_f(z)[\zeta, \zeta'] = \int_0^1 g(z)[\nabla_s\zeta, \nabla_s\zeta']\mathrm{d}s - \int_0^1 g(z)[R(\zeta, \dot{z})\dot{z}, \zeta']\, ds\,. \tag{16}$$

Now the linear operator $f''(z)$ associated with $H_f(z)$ is a Fredholm operator of index 0 on the tangent space $T_z\Omega^{1,2}(p, q; \mathcal{M})$. Indeed, we have that $H_f(z) = a_0(z) + k(z)$, where

$$a_0(z) = \int_0^1 g(z)[\nabla_s\zeta, \nabla_s\zeta']\, ds$$

is nondegenerate and

$$k(z) = -\int_0^1 g(z)[R(\zeta, \dot{z})\dot{z}, \zeta']\, ds$$

defines a compact linear operator on $T_z\Omega^{1,2}(p, q; \mathcal{M})$. This last fact is essentially a consequence of the compact embedding of the Sobolev space of functions defined on an interval into the space of continuous functions equipped with uniform convergence topology, or equivalently, it is a consequence of the classical Ascoli–Arzelà Theorem in Functional Analysis. So we can define the relative index and the augmented relative index for a geodesic.

Definition 7. *Let $(\mathcal{M}, g)$ be a semi-Riemannian manifold and let $z\colon [0,1] \to \mathcal{M}$ be a geodesic joining $p = z(0)$ and $q = z(1)$. The* relative index $j(z)$ *of the geodesic z is defined by setting*

$$j(z) = j(H_f(z), a_0(z))\ , \tag{17}$$

while the augmented relative index $j^*(z)$ *is defined by setting*

$$j^*(z) = j(z) + \dim\big(\ker H_f''(z)\big)\ .$$

Notice that $j(z) = j^*(z)$ if and only if p and q are nonconjugate along z. The index of a semi-Riemannian geodesic z is a relative integer $j(z) \in \mathbf{Z}$, so it could be negative. If the metric g is Riemannian, then the bilinear form $a_0(z)$ is positive definite and so the relative index $j(z)$ reduces to the

classical Morse index $m(z, f)$ of the geodesic. If the metric is Lorentzian, the spectral properties of $f''(z)$ are only partially known. If z is a *causal geodesic*, then $j(z) \in \mathbf{N}$, see [4, 8]. It would be interesting to classify the (spacelike) geodesics with negative index. It is very interesting to note that if $(\mathcal{M}, g, Y)$ is a stationary spacetime, then the index $j(z)$ of any geodesic for $(\mathcal{M}, g, Y)$ is nonnegative and $j(z) = m(z, J)$, where J is the restriction of the action integral to the natural constraint.

Actually there are no results on Morse relations for geodesics in nonstationary Lorentzian manifolds, but only on the Morse inequalities for geodesics joining two nonconjugate points in an orthogonally splitting one, see [2].

Theorem 16. *Let $(\mathcal{M}, g)$ be an orthogonally splitting Lorentzian manifold satisfying assumptions A_1)–A_5) and let p and q be two nonconjugate points of $\mathcal{M}$. Moreover, for any $k \in \mathbf{N}$, denote by $\mathcal{G}(p, q; k)$ the set of geodesics z for the metric g, joining p and q and such that the relative index $j(z)$ of z is equal to k. Then, for any $k \in \mathbf{N}$ and for any field $\mathcal{K}$ we have*

$$\sharp\mathcal{G}(p, q; k) \geq \beta_k(\Omega(\mathcal{M}); \mathcal{K}))\,, \tag{18}$$

where $\beta_k(\Omega(\mathcal{M}); \mathcal{K})$ is the k-th Betti number of the based loop space $\Omega(\mathcal{M}; \mathcal{K})$ with respect to the field $\mathcal{K}$.

Notice that under the assumptions of the previous theorem, whenever the manifold $\mathcal{M}$ is noncontractible to a point, there exist infinitely many geodesics joining the points p and q. The topological properties of the based loop space $\Omega(\mathcal{M})$ (cf. [22]) allows to estimate the relative index on a sequence of such geodesics. Indeed, it follows that a sequence of geodesics $(z_m)_{m\in\mathbf{N}}$ joining p and q exists such that $j(z_m) \to +\infty$ as $m \to +\infty$, see [2].

The proof of the Morse inequalities is based on a detailed analysis of the Morse theoretical properties of the family of functionals approximating the functional f introduced in [7], satisfying the Palais–Smale condition and having a finite dimensional saddle geometry. We refer to [2] for other details.

The Morse inequalities allow to estimate the number of geodesics having *nonnegative* index in terms of the singular homology groups of the based loop space. If one wants to give some estimates on the number of geodesics with *negative* index, classical homological or cohomological theories (as singular homology or singular cohomology) do not work. From the variational point of view, the classical theories are not affected by attaching an infinite dimensional cell, because the infinite dimensional unit sphere is contractible. The study of homology or cohomology theories for strongly indefinite functionals is an active field of interaction between Algebraic Topology and Critical Point Theory (see [1, 65] for some recent result). However, Morse relations for geodesics joining two nonconjugate points in a nonstationary Lorentzian manifold is a completely open problem.

6 Results on Manifolds with Boundary

Many physically relevant spacetimes exhibit a content of noncompleteness due to the presence of a boundary, a horizon or a singularity. We present now some results on the geodesic connectedness of some of these spacetimes admitting a horizon or a boundary. The proof of these results is essentially based on the following notion of convexity.

Definition 8. *Let $(\mathcal{M}, g)$ be a semi-Riemannian manifold and let $\mathcal{N}$ be an open subset of $\mathcal{M}$ with topological boundary $\partial\mathcal{N}$; we say that the open set $\mathcal{N}$ has a* convex boundary *$\partial\mathcal{N}$ if the following property holds: If a curve $z\colon [0,1] \to \mathcal{M}$ is a geodesic for the metric g such that $z([0,1]) \subset \mathcal{N} \cup \partial\mathcal{N}$ and $z(0)$, $z(1) \in \mathcal{N}$, then $z([0,1]) \subset \mathcal{N}$.*

The notion of an open set with convex boundary is different from the notion of a convex neighborhood introduced in many texts on differential geometry, being a global notion rather than a local one. If the boundary $\partial\mathcal{N}$ is a smooth submanifold of $\mathcal{M}$, the convexity can be determined by studying the restriction to the tangent bundle of $\partial\mathcal{N}$ of the Hessian of a smooth function representing a sort of distance from $\partial\mathcal{N}$, see [51].

Proposition 1. *Let $\mathcal{N}$ be an open subset of a semi-Riemannian manifold $(\mathcal{M}, g)$ with a smooth boundary $\partial\mathcal{N}$ and let $\Phi\colon \mathcal{M} \to \mathbf{R}$ be a smooth function such that $\Phi^{-1}(0) = \partial\mathcal{N}$, $\Phi^{-1}(]0, +\infty[) = \mathcal{N}$ and $\nabla\Phi(z) \neq 0$, for any $z \in \partial\mathcal{N}$.*

Then the open subset $\mathcal{N}$ has a convex boundary if and only if for any $z \in \partial\mathcal{N}$ and for any $v \in T_z\partial\mathcal{N}$, it is $H_\Phi(z)[v, v] \leq 0$, where $H_\Phi(z)$ denotes the Hessian operator of Φ at the point z with respect to the metric g.

If the boundary $\partial\mathcal{N}$ of the open subset $\mathcal{N}$ is *nonsmooth* (as it happens in many physically interesting spacetimes), the characterization of the convexity of $\partial\mathcal{N}$ is more delicate and requires careful estimates of H_Φ in a neighborhood of $\partial\mathcal{N}$ in $\mathcal{N}$.

Many global results have been obtained, using variational or topological methods, on geodesics joining two points in stationary or orthogonally splitting Lorentzian manifolds with boundary. We present here only the applications to physical spacetimes of General Relativity.

- *Schwarzschild spacetime.* Any pair of points outside the *event horizon* is joined by infinitely many distinct geodesics with images contained in the same region. Moreover the Morse relations and the Morse inequalities hold for any pair of nonconjugate points outside the event horizon, see [6, 35]. These results have been obtained by viewing the Schwarzschild spacetime outside the event horizon as an open subset, with nonsmooth boundary, of the Kruskal spacetime.
- *Reissner–Nordström spacetime.* The results stated for the Schwarzschild spacetime hold also for the static *external region* of the Reissner–Nordström spacetime with two event horizons. Moreover the existence of infinitely many geodesics joining two arbitrary points in the *intermediate*

region between the event horizons has been proved in [33]. If the points are nonconjugate, only the Morse inequalities actually hold. Notice that the Reissner–Nordström spacetime between the event horizons is orthogonally splitting but nonstatic.
– *Kerr spacetime*. The Kerr spacetime outside the *stationary limit surface* is geodesically connected, see [25].

The proof of the geodesic connectedness for the Kerr spacetime is based on a topological argument involving the topological degree defined for a suitable map. The proofs of the other results are variational in nature, using a perturbation argument for the action integral over the curves having support very close to the boundary of the open set considered. A priori estimates and a limit process for the critical points of these approximating functionals allow to prove the existence of a critical point of the action integral. For a different approach, using techniques of nonsmooth analysis, see [21].

7 Other Directions

In this paper we have presented some global properties of Lorentzian manifolds obtained by using variational methods. We have focused our attention on the geodesic connectedness of Lorentzian manifolds. Moreover, multiplicity results and Morse theoretical properties of such geodesics have been stated. These results have been obtained by exploiting the variational properties of the action integral on a Lorentzian manifold. In Lorentzian geometry, the action integral is strongly indefinite and its stationary points are infinite dimensional saddle points. Stationary points of the action integral cannot be found by minimization and more involved arguments based on critical point theorems for unbounded functionals are needed to solve the problem. On the other hand, many meaningful counter-examples to the geodesic connectedness are known in Lorentzian geometry.

Other interesting variational problems, such as the existence of closed geodesics of any causal character, or more generally, the existence of spatially closed geodesics, have not been considered. For results in these directions, see for instance [17, 29, 30, 45, 49, 66] and the references therein.

The Morse Index Theorem in Riemannian geometry states that the Morse index of a Riemannian geodesic z, that is the number of negative eigenvalues of the Hessian operator $f''(z)$, is equal to the number of conjugate points along the geodesic, counted with their multiplicity. The statement of this beautiful result cannot be immediately extended to geodesics on Lorentzian manifolds, because the Morse Index of a Lorentzian geodesic is equal to $+\infty$. The Morse Index Theorem was extended by Karen Uhlenbeck to timelike geodesics [67, 4], and by John Beem and Paul Ehrlich to lightlike ones [4]. The number of conjugate points along a timelike geodesic z is finite and it is equal to the number of negative eigenvalues of the restriction of the Hessian

operator $f''(z)$ to the subspace $\dot{z}^{\perp}$ of the tangent space $T_z\Omega^{1,2}(p,q;\mathcal{M})$ consisting of vector fields along z pointwise orthogonal to the tangent field $\dot{z}$. It is not difficult to show that the index of the restriction of $f''(z)$ to $\dot{z}^{\perp}$ is finite. In a certain sense, we have again deleted the time direction which is responsible for the indefiniteness of the geodesic problem in a Lorentzian manifold. The proof of the Morse Index Theorem for lightlike geodesics is delicate and involves more subtle reductions on the tangent space $T_z\Omega^{1,2}(p,q;\mathcal{M})$ than in the timelike case. The Morse Index Theorem for lightlike geodesics can be proved also in the context of relativistic optics and extensions to General Relativity of the Fermat principle, viewing light rays as critical points of an arrival time functional, see [37, 41, 58] for a proof. The Morse Index Theorem for spacelike geodesics has a totally different nature. First of all, it has been observed by Adam Helfer [47] that there exist a Lorentzian manifold and a spacelike geodesic admitting a continuum of conjugate points, see also [53, 61]. However, introducing a suitable nondegeneration assumption, it can be defined a spectral index for a spacelike geodesic which is equal to a relative integer, and it is equal to a geometric index for the geodesic, called the *Maslov index*, which is not given by simply counting conjugate points with their multiplicity but requires more complicated evaluations for a conjugate point, see [42, 53]. In general the Maslov index of a geodesic is a relative integer, too. For any spacelike geodesic z on a stationary Lorentzian manifold, the nondegeneration assumption holds and the spectral index is equal to the Morse index $m(z,J)$ of z as a critical point of the restriction J of the action integral to the natural constraint. The Morse Index Theorem claims that the spectral index $m(z,J)$ equals the Maslov index of z, so for spacelike geodesics on stationary Lorentzian manifolds, the Maslov index is a nonnegative number. For extensions to spacelike geodesics on nonstationary Lorentzian manifolds see [60].

Recently, some results have been obtained on the existence and multiplicity of solutions of the relativistic Lorentz force equation, and in particular for timelike solutions. The relativistic Lorentz force equation describes the motion of a charged massive test particle under the action of a gravitational field and an external electromagnetic field, see [68]. For results on existence and multiplicity of timelike solutions see [15, 16]. Also for the Lorentz force equation there are some open problems. In particular, we mention one of the most interesting problems about timelike solutions: it is still not clear if there are only finitely many conjugate points in a compact interval of a timelike solution of the relativistic Lorentz force equation, as for timelike geodesics, see also [18].

Finally, we mention that the problem of the geodesic connectedness of semi-Riemannian manifolds is still essentially open, except for generalizations of the results presented in Sects. 4 and 5 on stationary and orthogonally splitting Lorentzian manifolds.

References

1. A. Abbondandolo: *Morse Theory for Hamiltonian systems*, CRC Research Notes in Mathematics **425** (Chapman and Hall, London 2001) 54, 70, 72
2. A. Abbondandolo, V. Benci, D. Fortunato, A. Masiello: Math. Res. Lett. **10**, 435 (2003) 70, 72
3. D.E. Allison, B. Ünal: J. Geom. Phys. **46**, 193 (2003) 67
4. J.K. Beem, P. E. Ehrlich, K. L. Easley: *Global Lorentzian Geometry*, 2nd edn (Marcel Dekker, New York 1996) 51, 52, 72, 74
5. V. Benci, D. Fortunato, F. Giannoni: Ann. Inst. H. Poincaré, Analyse Nonlinéaire **8**, 79 (1991) 64, 65
6. V. Benci, D. Fortunato, F. Giannoni: Ann. Sc. Norm. Sup. (IV) **XIX**, 255 (1992) 73
7. V. Benci, D. Fortunato, A. Masiello: Math. Z. **217**, 73 (1994) 68, 72
8. V. Benci, F. Giannoni, A. Masiello: J. Geom. Phys. **27**, 267 (1998) 72
9. V. Benci, A. Masiello: Math. Ann. **293**, 433 (1992) 67
10. A.N. Bernal, M. Sánchez: Comm. Math. Phys. **243**, 461 (2003) 68
11. R. Bott: Bull. Am. Math. Soc **7**, 331 (1982) 60, 61
12. E. Calabi, L. Markus: Ann. Math. **75**, 63 (1962) 52
13. A.M. Candela, F. Giannoni, A. Masiello: J. Diff. Eq. **155**, 203 (1999) 61, 70
14. A.M. Candela, M. Sánchez: Diff. Geom. Appl. **12**, 105 (2000) 66
15. E. Caponio, A. Masiello: Class. Quantum Grav. **19**, 2229 (2002) 75
16. E. Caponio, A. Masiello: J. Math. Phys. **45**, 4134 (2004) 75
17. E. Caponio, A. Masiello, P. Piccione: Math. Z. **244**, 457 (2003) 66, 74
18. E. Caponio, A. Masiello, P. Piccione: Manuscr. Math. **113**, 471 (2004) 75
19. K.C. Chang: *Infinite Dimensional Morse Theory and Multiple Solutions Problems* (Birkhäuser, Boston 1993) 54, 70
20. C. Conley, E. Zehnder: Comm. Pure Appl. Math. **37**, 207 (1984) 70
21. M. Degiovanni, A. Giacomini: Nonlinear Anal. T. M. A. **47**, 5041 (2001) 74
22. E. Fadell, S. Husseini: Nonlinear Anal. T. M. A **17**, 1153 (1991) 62, 72
23. E. Fadell, S. Husseini: Rend. Sem. Mat. Fis. Univ. Milano, **LXIV**, 99 (1994) 61, 70
24. J.L. Flores, M. Sánchez: J. Geom. Phys. **36**, 285 (2000) 68, 69
25. J.L. Flores, M. Sánchez: J. Math. Phys. **43**, 4861 (2002) 74
26. J.L. Flores, M. Sánchez: this volume 66
27. D. Fortunato, F. Giannoni, A. Masiello: J. Geom. Phys. **15**, 159 (1995) 52
28. G. Fournier and M. Willem: Relative category and the calculus of variations. In *Variational Problems*, ed by H. Berestycki, J.M. Coron, I. Ekeland (Birkhäuser, Basel 1990) pp 95-104 61
29. G.J. Galloway: Trans. Am. Math. Soc. **285**, 379 (1984) 74
30. G.J. Galloway: Proc. Am. Math. Soc. **98**, 119 (1986) 74
31. R. Geroch: J. Math. Phys. **11**, 437 (1970) 68
32. F. Giannoni, A. Masiello: J. Funct. Anal. **101**, 340 (1991) 64, 65
33. F. Giannoni, A. Masiello: Manuscr. Math. **78**, 381 (1993) 74
34. F. Giannoni, A. Masiello: Ann. Inst. H. Poincaré, Analyse Nonlinéaire **12**, 27 (1995) 52, 61, 70
35. F. Giannoni, A. Masiello: Top. Meth. Nonlinear Anal. **6**, 1 (1995) 67, 73
36. F. Giannoni, A. Masiello, P. Piccione: Comm. Math. Phys. **187**, 1375 (1997) 62
37. F. Giannoni, A. Masiello, P. Piccione: Ann. Inst. H. Poincaré, Phys. Theor. **69**, 359 (1998) 62, 75

38. F. Giannoni, A. Masiello, P. Piccione: Class. Quantum Grav. **69**, 731 (1999) 62
39. F. Giannoni, A. Masiello, P. Piccione: J. Geom. Phys. **35**, 1 (2000) 52
40. F. Giannoni, A. Masiello, P. Piccione: Gen. Rel. Grav. **33**, 491 (2001) 62
41. F. Giannoni, A. Masiello, P. Piccione: J. Math. Phys. **43**, 563 (2002) 52, 75
42. F. Giannoni, A. Masiello, P. Piccione, D. Tausk: Asian J. Math. **3**, 441 (2001) 67, 75
43. F. Giannoni, P. Piccione: Comm. Anal. Geom. **7**, 157 (1999) 64, 65
44. F. Giannoni, P. Piccione, R. Sampalmieri: J. Math. Anal. Appl. **252**, 444 (2000) 66, 67
45. M. Guediri: Math. Z. **239**, 277 (2002) 74
46. S.W. Hawking, G.F.R. Ellis: *The Large Scale Structure of Spacetime*, (Cambridge University Press, Cambridge 1973) 51, 64, 68
47. A. Helfer: Pac. J. Math. **164**, 321 (1994) 75
48. W. Klingenberg: *Riemannian Geometry*, 2nd edition (W. De Gruyter, Berlin 1995) 53, 61
49. A. Masiello: J. Diff. Eq. **104**, 48 (1993) 74
50. A. Masiello: Ann. Mat. Pura Appl. (IV) **CLXVII**, 299 (1994) 69
51. A. Masiello: *Variational Methods in Lorentzian Geometry*, Pitman Research Notes in Mathematics **309** (Longman, London 1994) 54, 55, 73
52. J. Mawhin, M. Willem: *Critical Point Theory and Hamiltonian Systems*, (Springer, Berlin 1989) 54, 57, 58, 60, 61
53. F. Mercuri, P. Piccione, D. Tausk: Pac. J. Math. **206**, 375 (2002) 75
54. J. Milnor: *Morse Theory*, Ann. of Math. Studies **51** (Princeton Univ. Press, Princeton 1963) 56, 61
55. B. O'Neill: *Semi-Riemannian Geometry with Applications to Relativity*, (Academic Press, New York 1983) 51, 52, 63, 68
56. R. Palais: Topology **2**, 299 (1963) 59, 61
57. R. Penrose: *Techniques of Differential Topology in Relativity*, Regional Conference Series in Applied Math. **7** (Society for Industrial and Applied Mathematics, Philadelphia 1972) 66
58. V. Perlick: *Ray Optics, Fermat's Principle, and Applications to General Relativity*, Lecture Notes in Physics **m61** (Springer, Heidelberg 2000) 75
59. V. Perlick: Living Rev. Relativity **7** (2004), 9. `http://www.livingreviews.org/lrr-2004-9` 52
60. P. Piccione, D. Tausk: Proc. London Math. Soc. (3) **83**, 351 (2001) 75
61. P. Piccione, D. Tausk: Comm. Anal. Geom. **11**, 33 (2003) 75
62. J.P. Serre: Ann. Math. **54**, 425 (1951) 62
63. E.H. Spanier: *Algebraic Topology* (McGraw Hill, New York 1966) 59
64. M. Struwe: *Variational Methods*, 3rd edn (Springer, Heidelberg 2000) 54, 57, 61
65. A. Szulkin: Math. Z. **209**, 375 (1992) 72
66. F.J. Tipler: Proc. Am. Math. Soc. **76**, 145 (1979) 74
67. K. Uhlenbeck: Topology **14**, 69 (1975) 52, 74
68. R. Wald: *General Relativity* (University of Chicago Press, Chicago 1984) 51, 63, 75

On the Geometry of pp-Wave Type Spacetimes

José L. Flores[1,2] and Miguel Sánchez[3]

[1] Department of Mathematics, Stony Brook University, Stony Brook, NY 11794-3651, USA
floresj@math.sunysb.edu
[2] Permanent address: Departamento de Álgebra, Geometría y Topología, Universidad de Málaga, Campus Teatinos, 29071 Málaga, Spain
floresj@agt.cie.uma.es
[3] Departamento de Geometría y Topología, Facultad de Ciencias, Universidad de Granada, Avenida Fuentenueva s/n, E-18071 Granada, Spain
sanchezm@ugr.es

Abstract. Global geometric properties of product manifolds $\mathcal{M} = M \times \mathbb{R}^2$, endowed with a metric type $\langle \cdot, \cdot \rangle = \langle \cdot, \cdot \rangle_R + 2dudv + H(x,u)du^2$ (where $\langle \cdot, \cdot \rangle_R$ is a Riemannian metric on M and $H : M \times \mathbb{R} \to \mathbb{R}$ a function), which generalize classical plane waves, are revisited. Our study covers causality (causal ladder, non-existence of horizons), geodesic completeness, geodesic connectedness and existence of conjugate points. Appropriate mathematical tools for each problem are emphasized and the necessity to improve several Riemannian (positive definite) results is claimed.

The behaviour of $H(x,u)$ for large spatial component x becomes essential, with a spatial quadratic behaviour being critical for many geometrical properties. In particular, when M is complete, if $-H(x,u)$ is spatially subquadratic, the spacetime becomes globally hyperbolic and geodesically connected. But if a quadratic behaviour is allowed (as happens in plane waves), then both global hyperbolicity and geodesic connectedness may be lost.

From the viewpoint of classical general relativity, the properties which remain true under generic hypotheses on $\mathcal{M}$ (as subquadraticity for H) become meaningful. Natural assumptions on the wave – finiteness or asymptotic flatness of the front – imply the spatial subquadratic behaviour of $|H(x,u)|$ and, thus, strong results for the geometry of the wave. These results do not always hold for plane waves, which appear as an idealized non-generic limit case.

1 Introduction

Among the reasons which contribute to the recent interest in pp-wave type spacetimes, we mention, on the one hand, classical geometrical properties and, on the other, applications to string theory[1]. About the former, pp-waves

[1] Of course, there is also another very influential reason: the possibility of direct detection of gravitational waves. Hulse and Taylor were awarded the Nobel prize in 1993 for the discovery, in the 1970s, of indirect evidence of their existence – a binary

J.L. Flores and M. Sánchez: *On the Geometry of pp–Wave Type Spacetimes*, Lect. Notes Phys. **692**, 79–98 (2006)
www.springerlink.com

spacetimes, and especially plane waves, [12, 23, 35] have curious and intriguing properties, which led to questions still open or only recently solved. The well-known Penrose limit [39] (see also [9, 10]) associates to every spacetime and every choice of an (unparametrized) lightlike geodesic a plane-wave metric. Penrose [37] also emphasized that, in spite of being geodesically complete, plane waves are not globally hyperbolic (see Sect. 2 for definitions). Ehlers and Kundt [19] conjectured that gravitational plane waves are the only complete gravitational pp-waves. As we will see, by now the lack of global hyperbolicity is well understood, but the Ehlers-Kundt conjecture still remains open. Applications to string theory have highlights such as: (a) gravitational pp-waves are relevant spacetimes with vanishing scalar invariants (VSI, see [17, 40] for a classification), and such spacetimes yield exact backgrounds for string theory (vanishing of α' corrections, see [2, 28]), (b) Berenstein, Maldacena and Nastase [5] have recently proposed an influential solvable model for string theory by taking the Penrose limit in $\mathrm{AdS}_5 \times S^5$ spacetimes, or (c) after realizing that Gödel–like universes can be supersymmetrically embedded into string theory, it was realized and emphasized that these solutions were T-dual to compactified plane-wave backgrounds [11, 26, 33].

The necessity to better understand the geometry of waves and their potential applications to string theory justifies to study pp-waves from a wider perspective, where new mathematical tools appear naturally. The authors, in collaboration with A.M. Candela [15], considered the following class of spacetimes $(\mathcal{M}, \langle\cdot,\cdot\rangle)$, called *plane-fronted waves* (PFW) or Mp-*waves*, see Sect. 2, which essentially include the classical pp-waves and, thus, plane waves:

$$\begin{aligned} \mathcal{M} &= M \times \mathbb{R}^2 \\ \langle\cdot,\cdot\rangle &= \langle\cdot,\cdot\rangle_R + 2\, du\, dv + H(x,u)\, du^2\ . \end{aligned} \tag{1}$$

Here $(M, \langle\cdot,\cdot\rangle_R)$ is any smooth Riemannian (C^∞, positive-definite, connected) n-manifold, the variables (v,u) are the natural coordinates of $\mathbb{R}^2$ and the smooth scalar field $H : M \times \mathbb{R} \to \mathbb{R}$ is not identically 0.

Our initial motivation to study such metrics came from some work by two contributors to this meeting, R. Penrose and P.E. Ehrlich. Penrose [37] showed that, even though plane waves are strongly causal, they are not globally hyperbolic. Moreover, they present a property of focusing lightlike geodesics which forbids not only global hyperbolicity but also the possibility to embed them isometrically in higher-dimensional semi-Euclidean spaces

---

system loses an exact amount of rotational energy which can be interpreted only as originating from the radiation of gravitational waves. Nowadays, experimentalists look for direct evidence, and a generation of large scale interferometers is close to be operative in various places on the Earth (VIRGO, LIGO, GEO300, TAMA300 ...) and even in space (LISA). Although experimentalists' problems are very different from the ones in this paper, if they succeed an excellent stimulus on waves for the whole relativistic community (and even for the curiosity of the general public) will be achieved.

(Fig. 1). This is a *remarkable* property of plane waves but, as he pointed out, it is also interesting to know "whether the somewhat strange properties of plane waves encountered here will be present for waves which approximate plane waves, but for which the spacetime is asymptotically flat, or asymptotically cosmological in some sense". From our viewpoint, this is a relevant question because the geometrical properties of an exact solution to Einstein's equation (as plane waves) are physically meaningful only if they are "stable" in some sense – surely, this is not fulfilled by a term as H in formula (2) below. Even more, in the setting of Penrose's *Strong Cosmic Censorship Hypothesis* [38], generic solutions to Einstein's equation with reasonable matter and behaviour at infinity must be globally hyperbolic. And, obviously, plane waves fail to be generic and well behaved at infinity because of the many symmetries of the term H (as well as the part $M = (\mathbb{R}^2, dx^2 + dy^2)$).

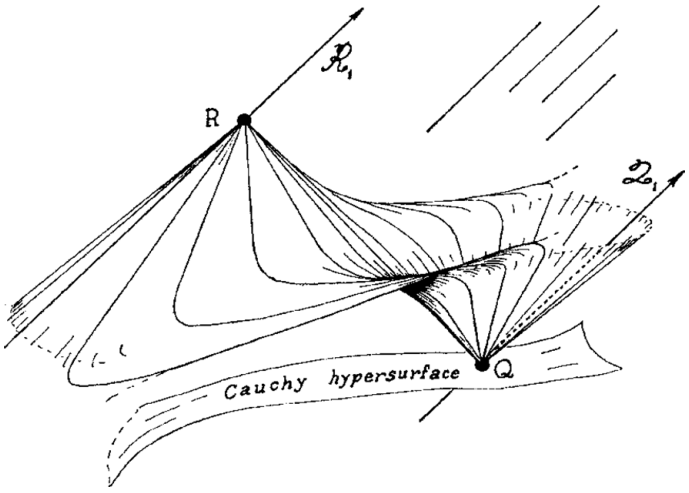

Fig. 1. Focusing of lightlike geodesics in plane waves; the picture is taken from [37]

Ehrlich and Emch, in a series of papers [20, 21, 22] (see also [4, Ch. 13]), carried out a detailed investigation of the behaviour of all the geodesics emanating from a (suitably chosen) point p in a gravitational plane wave. Then, they showed that gravitational plane waves are causally continuous but not causally simple, and characterized points necessarily connectable by geodesics (see Subsect. 5.1). However, again all of the study relies on the "non-generic nor stable" conditions of symmetry of the gravitational wave and on the very special form of $H(x, u)$: independence of the choice of the point p, explicit integrability of geodesic equations, identical equations for Killing fields, Jacobi fields and geodesics.

In this framework, our goals in [15, 24] were, essentially: (i) to introduce the class of reasonably generic waves (1), (ii) to justify that, for a

physically reasonable asymptotic behaviour of the wave, $|H(\cdot, u)|$ must be "subquadratic" (plane waves lie in the limiting quadratic case), and (iii) to show that, in this case, the geometry of the wave presents good global properties, including global hyperbolicity [24] and geodesic connectedness [15]. Even more, the instability of these geometric properties in the quadratic case leads to interesting questions in Riemannian Geometry, studied in [13].

In the present article, we explain the role of the mathematical tools introduced in [13, 15, 24] in relation to classical papers on waves such as [19, 37], and to later developments [29, 30, 31, 32, 33, 44]. For proofs we refer to the original articles or, in the case of more recent results, we provide sketches.

This paper is organised as follows:

In Sect. 2, some general properties of PFWs are explained, including questions related to curvature and the energy conditions. Remarkably, we justify that the behaviour of $|H(x, u)|$ for large x must be subquadratic if the wave is assumed to be finite or with fronts asymptotically flat in any reasonable sense. This becomes relevant from the viewpoint of classical general relativity, and the global geometrical properties of PFWs will depend dramatically on the possible quadratic behaviour of H or $-H$ (for any dimension $n \geq 1$ of M).

In Sect. 3, we show that the behaviour of all the causal curves can be essentially controlled in a PFW (the more accurate control for existence of causal geodesics is postponed to Sect. 5). In Subsect. 3.1 a detailed study of the causal hierarchy of PFWs is carried out. In particular, Penrose's abovementioned question is answered by showing that the causal hierarchy of plane waves is "unstable" or "critical": deviations in the superquadratic direction of $-H$ may transform them in non-distinguishing spacetimes, but deviations in the (more realistic) subquadratic direction yield global hyperbolicity. Subsequent results by Hubeny, Rangamani and Ross [32] are also discussed. In Subsect. 3.2 the criterion for the non-existence of horizons posed by Hubeny and Rangamani [30, 31] is explained, and a simple proof showing that it holds for any PFW is given.

In Sect. 4, geodesic completeness is studied. We claim that this problem is equivalent to a purely Riemannian problem (Theorem 2), which has been solved satisfactorily only for autonomous H, i.e., $H(x, u) \equiv H(x)$. The power of the known autonomous results (which yield completeness for at most quadratic $H(x)$, Theorem 3) is illustrated by comparison with the examples in [29]. Then, we claim the necessity to improve the non-autonomous ones. Moreover, the Ehlers–Kundt conjecture deserves a special discussion. Even though easily solvable under at most quadraticity for x (Theorem 4), it remains open in general.

In Sect. 5, the problems related to geodesic connectedness are studied. The key is to reduce the problem to a purely Riemannian problem, concretely, the classical variational problem of finding critical points for a Lagrangian type kinetic energy minus (time-dependent) potential energy. That is, solving this

classical problem becomes equivalent to solving the geodesic connectedness problem in PFWs. Remarkably, in order to obtain the optimal results on waves (extending Ehrlich-Emch's ones) we had to improve the known Riemannian results; in the Appendix, this Riemannian problem is explained. Finally, the existence of conjugate points is discussed, and reduced again to a purely Riemannian problem. Energy conditions tend to yield conjugate points for causal geodesics. But, in agreement with the other results of the present paper, the focusing property of lighlike geodesics in plane waves (Fig. 1) becomes highly non-generic.

2 General Properties of the Class of Waves

2.1 Definitions

Let us start with some simple properties of the metric (1). The assumed geometrical background material can be found in well-known books such as [4, 27, 36]. Following [36], vector 0 will be regarded as spacelike instead of lightlike.

The vector field ∂_v is absolutely parallel (i.e., covariantly constant) and lightlike, and the time-orientation will be chosen to make it past-directed. Thus, for any future-directed causal curve $z(s) = (x(s), v(s), u(s))$,

$$\dot{u}(s) = \langle \dot{z}(s), \partial_v \rangle \geq 0 \, ,$$

and the inequality is strict if $z(s)$ is timelike. As $\mathrm{grad} u = \partial_v$, coordinate $u : \mathcal{M} \to \mathbb{R}$ plays the role of a "quasi-time" function [4, Def. 13.4], i.e., its gradient is everywhere causal and any causal segment γ with $u \circ \gamma$ constant (necessarily a lightlike pregeodesic without conjugate points) is injective. In particular, the spacetime is causal (see also Sect. 3.1). The hypersurfaces $u \equiv$ constant are degenerate, with the kernel of the metric spanned by ∂_v. The hypersurfaces (non-degenerate n-submanifolds) of these degenerate hypersurfaces which are transverse to ∂_v, must be isometric to open subsets of M. The *fronts* of the wave (1) will be defined as the (whole) n-submanifolds at constant u, v.

According to Ehlers and Kundt [19] (see also [8]) a vacuum spacetime is a plane-fronted gravitational wave if it contains a shearfree geodesic lightlike vector field V, and admits "plane waves" – spacelike (two-)surfaces orthogonal to V. The best known subclass of these waves are the (gravitational) "plane-fronted waves with parallel rays" or pp-waves, which are characterized by the condition that V is covariantly constant, $\nabla V = 0$. Ehlers and Kundt gave several characterizations of these waves in coordinates, and they obtained that (at least locally) the metric can be written as in (1) with $M = \mathbb{R}^2$. Nowadays, pp-wave means any spacetime which admits a covariantly constant lightlike vector field [45, p. 383]. Even though, in general, their

fronts may be "non-plane", this happens in the most relevant cases (four dimensional spacetimes which are either vacuum, or solutions to Einstein-Maxwell equations, or pure radiation fields), where expression (1) holds with $M = \mathbb{R}^2$. In what follows, we will use the name "pp-waves" to denote the classical spacetimes (1) with $M = \mathbb{R}^2$. The pp-wave is gravitational (i.e., vacuum, see Subsect. 2.2) if and only if the "spatial" (transverse) Laplacian $\Delta_x H(x,u)$ vanishes. Plane waves constitute the (highly symmetric) subclass of pp-waves with H exactly quadratic in x for appropriate global coordinates on each front; that is, when we can assume:

$$H(x,u) = (x^1, x^2) \begin{pmatrix} f_1(u) & g(u) \\ g(u) & -f_2(u) \end{pmatrix} \begin{pmatrix} x^1 \\ x^2 \end{pmatrix} \tag{2}$$

where f_1, f_2, g are arbitrary (smooth) functions. When $f_1 \equiv f_2$, the plane wave is gravitational, and there are other well-known subclasses ("sandwich plane wave" if f_1, f_2, g have compact support; "purely electromagnetic plane wave" if $f_1 \equiv -f_2$, $g \equiv 0$, etc.)

Recall that, in our type of metrics (1), no restriction on the Riemannian part $(M, \langle \cdot, \cdot \rangle_R)$ is imposed. This seems convenient for different reasons as, for example: (i) the generality in the dimension n, for applications to strings, (ii) the generality in the topology, for discussions of horizons, or (iii) the generality in the metric, to obtain "generic results", not crucially dependent on very special particular properties of the metric. In this context, a name such as "M-fronted wave with parallel rays" (Mp-wave) seems natural for our spacetimes (1). Nevertheless, we will maintain the name PFW (plane-fronted wave) in agreement with previous references [15, 24] and the nomenclature in [4], but with no further pretension.

2.2 Curvature and Matter

Fixing some local coordinates $x^1, \ldots, x^n$ for the Riemannian part M, it is straightforward to compute the Christoffel symbols of $\langle \cdot, \cdot \rangle$ and, thus, to relate the Levi–Civita connections ∇, ∇^R for $\mathcal{M}$ and M, respectively (see [15]). We remark the following facts:

- M is totally geodesic, i.e., $\nabla_{\partial_i} \partial_j = \nabla^R_{\partial_i} \partial_j$, $i, j = 1, \ldots, n$.
- $2\,\nabla_{\partial_u} \partial_u = -\mathrm{grad}_x H(x,u) + \partial_u H(x,u) \partial_v$; $\quad 2\nabla_{\partial i} \partial_u = \partial_i H(x,u) \partial_v$; thus, the curvature tensor satisfies

$$-\,2\,R(\cdot, \partial_u, \partial_u, \cdot) = \mathrm{Hess}_x H(x,u)(\cdot, \cdot)\,. \tag{3}$$

  Here $\mathrm{grad}_x H$ and $\mathrm{Hess}_x H$ denote the spatial (or "transverse") gradient and Hessian of H, respectively.
- The Ricci tensors of $\mathcal{M}$ and M satisfy

$$\mathrm{Ric} = \sum_{i,j=1}^{n} R^{(R)}_{ij} dx^i \otimes dx^j - \frac{1}{2} \Delta_x H\, du \otimes du\,.$$

Thus, Ric is zero if and only if both the Riemannian Ricci tensor $\mathrm{Ric}^{(R)}$ and the spatial Laplacian $\Delta_x H$ vanish.

From the last item, it is easy to check that the timelike convergence condition holds if and only if

$$\mathrm{Ric}^{(R)}(\xi,\xi) \geq 0, \quad \Delta_x H \leq 0, \quad \text{for all } \ x \in M, \ \ \xi \in T_x M \ .$$

Even more, in dimension 4 all the energy conditions are equivalent and easily characterized [24, Proposition 5.1]:

Proposition 1. *Let $\mathcal{M} = M \times \mathbb{R}^2$ be a 4-dimensional PFW, and let $K(x)$ be the (Gauss) curvature of the 2-manifold M. The following conditions are equivalent:*

(A) *The strong energy condition ($Ric(\xi,\xi) \geq 0$ for all timelike ξ).*
(B) *The weak energy condition ($T(\xi,\xi) \geq 0$ for all timelike ξ).*
(C) *The dominant energy condition ($-T^a_b \xi^b$ is either 0 or causal and future-pointing, for all future-pointing timelike $\xi \equiv \xi^b$).*
(D) *Both inequalities:*

$$K(x) \geq 0, \quad \Delta_x H(x,u) \leq 0, \qquad \forall (x,u) \in M \times \mathbb{R} \ .$$

2.3 Finiteness of the Wave and Decay of H at Infinity

Now, let us discuss *minimal necessary* assumptions which must be satisfied by a PFW to be "finite" or "asymptotically vanishing" in any reasonable sense. In principle, one could think that M should be asymptotically flat, but we will not impose this strong condition a priori (i.e., non-trivial fronts at a "cosmological scale" are admitted). In any case, it would not be too relevant for our problem: plane waves have flat fronts and are not finite by any means.

As we have said, all the scalar curvature invariants of a gravitational pp-wave vanish. Thus, instead of such scalars, we will focus on the spatial Hessian $\mathrm{Hess}_x H$. In the case of plane waves, $\mathrm{Hess}_x H$ is essentially the matrix in (2) – it can be viewed as the *transverse frequency matrix* of the lightlike geodesic deviation [37]. By equality (3), $\mathrm{Hess}_x H$ is related to the most "characteristic" curvatures of the wave; these curvatures – taken along a lightlike geodesic – admit an intrinsic interpretation in terms of the Penrose limit (see [9], especially the discussions around formulas (1.2), (2.13)). According to [30, 31], "to go arbitrarily far" in a pp-wave can be thought of as taking v, x large for each fixed u (see also Subsect. 3.2). Therefore, any sensible definition of finiteness or asymptotic vanishing of the PFW seems to imply that $\mathrm{Hess}_x H(x,u)$ must go (fast) to 0 for large x.

Rigourously, let $\lambda_i(x,u), i = 1,\ldots,n$, be the eigenvalues of $\mathrm{Hess}_x H(x,u)$, $d(\cdot,\cdot)$ the Riemannian distance on M and fix any $\bar{x} \in M$. From the above

discussion, if the wave vanishes asymptotically then $\lim_{d(x,\bar{x})\to\infty} \lambda_i(x,u)$ must vanish fast for each u. Therefore, putting $|\lambda(x,u)|$ equal to the maximum of the $|\lambda_i(x,u)|$'s, we can assume as definition of asymptotic vanishing for a PFW:

$$|\lambda(x,u)| \leq \frac{A(u)}{d(x,\bar{x})^{q(u)}} \tag{4}$$

for some continuous functions $A(u)$ and $q(u) > 0$.

Inequality (4) implies bounds for the spatial growth of $|H|$, as the next proposition shows. But, first, let us introduce the following definition. Let $V(x,u)$ be a continuous function $V : M \times \mathbb{R} \to \mathbb{R}$. We will say that $V(x,u)$ behaves *subquadratically at spatial infinity* if

$$V(x,u) \leq R_1(u) d^{p(u)}(x,\bar{x}) + R_2(u) \quad \forall (x,u) \in M \times \mathbb{R}\,,$$

for some continuous functions $R_1(u), R_2(u)(\geq 0)$ and $p(u) < 2$. If the last inequality is relaxed into $p(u) \leq 2, \forall u \in \mathbb{R}$ then $V(x,u)$ behaves *at most quadratically at spatial infinity*. Now, we can assert the following result (see [24, Proposition 5.3] for the idea of the proof – notice that the completeness of M is not necessary now, as any curve can be approximated by broken geodesics).

Proposition 2. *If the PFW vanishes asymptotically as in (4), then $|H(x,u)|$ behaves subquadratically at spatial infinity.*

The following must be emphasized:

1. The condition of asymptotic vanishing (4) implies subquadraticity for $|H(x,u)|$, but the converse is not true. In the remainder of this paper, we will use only this more general subquadratic condition or, even, only the subquadraticity (or at most quadraticity) of H or $-H$. So, the range of application of our results will be wider.
2. Of course, inequality (4) is compatible with the energy conditions. A simple explicit example can be constructed by taking $H(x^1,x^2,u) \equiv H_u(x^1)$, and, putting $x \equiv x^1$:

$$-\frac{A(u)}{|x|^{q(u)}} \leq \frac{d^2 H_u}{dx^2}(x) \leq 0$$

   for some $A(u), q(u) > 0$.
3. For plane waves, neither H nor $-H$ behaves subquadratically. In fact, the eigenvalues of $\mathrm{Hess}_x H(x,u)$ are independent of x, and the fronts of the wave are not "finite". This is a consequece of the idealized symmetries of the front waves. Nevertheless, $|H(x,u)|$ behaves at most quadratically at infinity and, thus, plane waves lie in the limiting quadratic situation.

3 Causality

3.1 Positions in the Causal Ladder

Recall first the causal hierarchy of spacetimes [4]:

$$\begin{gathered}\text{Globally hyperbolic} \Rightarrow \text{Causally simple} \Rightarrow \text{Causally continuous} \\ \Rightarrow \text{Stably causal} \Rightarrow \text{Strongly causal} \\ \Rightarrow \text{Distinguishing} \Rightarrow \text{Causal} \Rightarrow \text{Chronological}\end{gathered}$$

Roughly, a spacetime is causal if it does not contain closed causal curves, strongly causal if there are no "almost closed" causal curves and stably causal if, after opening slightly the light cones, the spacetime remains causal. It is commonly accepted that stable causality is equivalent to the existence of a *continuous* time function (see [4, 27], and also [43, Sect. 4]), but only recently has the existence of a *smooth* time function with nowhere lightlike gradient – i.e., a "temporal" function – been proven [7]. Globally hyperbolic spacetimes can be defined as the strongly causal ones with compact diamonds $J^+(p) \cap J^-(q)$ for any p, q. They were characterized by Geroch as those possessing a Cauchy hypersurface (which can be also chosen smooth and spacelike [6]). PFWs are always causal (Sect. 2) and the following result was proven in [24]:

Theorem 1. *Any PFW with M complete and $-H$ spatially subquadratic is globally hyperbolic.*

The following points must be emphasized:

1. The proof is carried out by showing strong causality and the compactness of the diamonds. From the technique, one can also check that, if $-H$ is at most quadratic at spatial infinity, then the spacetime is strongly causal (with no assumption on the completeness of M).
2. Hubeny, Rangamani and Ross [32] constructed explicitly a temporal function for plane waves. As the light cones of an at most quadratic pp-wave can be bounded by the cones of a plane wave, they claim that *any pp-wave with $-H$ at most quadratic at spatial infinity is stably causal*. (They also use the temporal function to study quotients of the wave by the isometry group generated by a spacelike Killing field, which may be stably causal or non-chronological, see also [33]).

   Recall from the Introduction that gravitational plane waves are causally continuous (the set valued maps $I^{\pm}$ are outer continuous) but not causally simple (because the causal future or past of a point may be non-closed).
3. If $-H(x, u)$ were not at most quadratic, then the spacetime may be even non-distinguishing (the chronological future or past of two distinct points are equal). In fact, a wide family of non-distinguishing examples with

$-H$ "arbitrarily close" to at most quadratic (and M complete) is constructed in [24, Proposition 2.1]; in these PFWs, the chronological futures $I^+(x, v, u)$ depend only on u. In particular, any pp-wave such that $-H$ behaves as $|x|^{2+\epsilon}, \epsilon > 0$ for large $|x|$ is non-distinguishing [24, Example 2.2].

Nevertheless, the spatially subquadratic or at most quadratic behaviour of $-H$ is not necessary for global hyperbolicity or strong/stable causality, as explicit counterexamples [24, Example 4.5] show (additional hypotheses must be assumed, see [32, Sect. 4]).

4. A curious phenomenon, suggested in [32, Sect. 4], is that the class of distinguishing but non-stably causal pp-waves (or even PFWs) might be empty. In this respect, our technique in [24] suggests that, if non-empty, the class would be scarcely significant.

The technique involved for Theorem 1 can be understood as follows. Any future-directed timelike curve α can be reparametrized by the quasi-time u: $\alpha(u) = (x(u), v(u), u), u \in [u_0, u_1]$. The proof is based on inequalities which relate the distance covered by $x(u)$ with the extreme points of $v(u)$. For fixed $\epsilon > 0$ and $0 < u_1 - u_0 \leq \epsilon$, $u \in [u_0, u_1]$:

$$\frac{1}{\epsilon^2} \int_{u_0}^{u} d^2(x(s), x(u_0)) ds \leq \int_{u_0}^{u} \langle \dot{x}(s), \dot{x}(s) \rangle_R ds$$

$$< 4 \left(R_2' \cdot (u - u_0) - (v(u) - v(u_0)) \right)$$

where the constant $R_2' = R_2'(u_0, \epsilon)$ is independent of $x(u_0)$ in the subquadratic case (in the finer proof of strong causality for the at most quadratic case, R_2' is allowed to depend on a compact subset where $x(u_0)$ lies, and $\epsilon > 0$ is not fixed a priori). Then, such an inequality is used:

- For the proof of strong causality, to show that, fixed a small neighborhood $\mathcal{N}$ of a point z_0 (which can be chosen "rectangular" in the coordinates v, u, i.e., $\mathcal{N} = N \times]\, v^-, v^+ [\times]\, u^-, u^+ [\,$), and any causal curve with extremes in this neighborhood, the restrictions on the extremes for $v(u), u$ $(v_0, v_1 \in]\, v^-, v^+ [\ , u_0, u_1 \in]\, u^-, u^+ [\,)$ also imply restrictions on the distance between $x(u_0), x(u_1)$. This forces the whole curve $x(u)$ to remain in a controlled neighborhood of N.
- For global hyperbolicity, to prove also that the projections of each diamond $J^+(p) \cap J^-(q) \subset M \times \mathbb{R}^2$ on each factor $M, \mathbb{R}^2$ are bounded for the natural (complete) Riemannian distances d on M and $du^2 + dv^2$ on $\mathbb{R}^2$. Therefore $J^+(p) \cap J^-(q)$ will be included in a compact subset, which in turn yields its compactness.

3.2 Causal Connectivity to Infinity and Horizons

Next, let us comment on the applicability of these techniques to the study of horizons in PFWs. The possible existence of horizons in gravitational

pp-waves and, in general, in vanishing scalar curvature invariant (VSI) spacetimes, have attracted interest recently, especially motivated by potential applications to string theory. Hubeny and Rangamani [30, 31] proposed a criterion for the existence of horizons in pp-waves, and they proved the *non-nexistence* of such horizons. In a more standard framework, Senovilla [44] proved the non-existence of closed trapped or nearly trapped surfaces (or submanifolds in any dimension) in VSI spacetimes. Next, we will give a simple proof of the non-existence of horizons, in the sense of Hubeny and Rangamani, for an arbitrary PFW.

Hubeny-Rangamani's criterion [30, Sects. 2.2, 4] can be reformulated as follows[2]: *A pp-wave spacetime (or, in general, any PFW) $\mathcal{M}$ does not admit an event horizon if and only if, given any points $z_0 = (x_0, v_0, u_0), z_1 = (x_1, v_1, u_1) \in \mathcal{M}$ with $u_0 < u_1$, there is $-v_\infty > -v_1$ such that a future-directed causal curve from z_0 to $z_\infty = (x_1, v_\infty, u_1)$ exists.* According to the authors, this criterion tries to formalize the intuitive idea that any point of the spacetime is visible to an observer who is "arbitrarily far". In fact, one may think of u_1 as being close to u_0 and of x_1 as being far[3] from x_0.

To check that this criterion is satisfied for any PFW, choose any curve α starting at z_0 parametrized by u, $\alpha(u) = (x(u), v(u), u), u \in [u_0, u_1]$ such that $x(u_1) = x_1$. Putting $E_\alpha(u) = \langle \dot\alpha(u), \dot\alpha(u)\rangle = \langle \dot x(u), \dot x(u)\rangle_R + 2\dot v(u) + H(x(u), u)$, the function $v(u)$ can be re-obtained from $E_\alpha(u)$ as:

$$v(u) - v_0 = \frac{1}{2}\int_{u_0}^{u} (E_\alpha(\bar u) - \langle \dot x(\bar u), \dot x(\bar u)\rangle_R - H(x(\bar u), \bar u))\, d\bar u, \quad \forall u \in [u_0, u_1]\ .$$

Choosing $E_\alpha < 0$ the curve α becomes timelike and future directed and, as $|E_\alpha|$ can be chosen arbitrarily big (and even constant, if preferred), the value of $-v(u_1)$ can be taken arbitrarily big, as required.

4 Geodesic Completeness

4.1 Generic Results

From the direct computation of Christoffel symbols of a PFW, it is straightforward to write the geodesic equations in local coordinates. Remarkably, the three geodesic equations for a curve $z(s) = (x(s), v(s), u(s))$, $s \in]\, a, b\,[$, can be solved in the following three steps [15, Proposition 3.1]:

(a) $u(s)$ is any affine function, $u(s) = u_0 + s\Delta u$, for some $\Delta u \in \mathbb{R}$.

[2]There is a change of sign for v in comparison to this reference because our convention for the metric uses $du\, dv$ instead of $-du\, dv$.

[3]Nevertheless, in our reformulation x_1 may also be "non-far" from x_0 (in fact, "to be far" would make no sense if M were compact or bounded). The criterion will hold in this strong sense.

(b) Then $x = x(s)$ is a solution on M of

$$D_s\dot{x} = -\mathrm{grad}_x V_\Delta(x(s), s) \quad \text{for all } s \in]\, a, b\,[\,,$$

where D_s denotes the covariant derivative and V_Δ is defined as[4] :

$$V_\Delta(x, s) = -\frac{(\Delta u)^2}{2}\, H(x, u_0 + s\Delta u)\,;$$

(c) Finally, with a fixed v_0 and an $s_0 \in]\, a, b\,[$, $v(s)$ can be computed from:

$$v(s) = v_0 + \frac{1}{2\Delta u}\int_{s_0}^{s} (E_z - \langle \dot{x}(\sigma), \dot{x}(\sigma)\rangle_R + 2V_\Delta(x(\sigma), \sigma))\, d\sigma$$

where $E_z = \langle \dot{z}(s), \dot{z}(s)\rangle$ is a constant (if $\Delta u = 0$ then $v = v(s)$ is also affine).

In particular, geodesic completeness is reduced, essentially, to the completeness of trajectories for (non-autonomous) potentials on M, and one can prove [15, Theorem 3.2]:

Theorem 2. *A PFW is geodesically complete if and only if the Riemannian manifold M is complete and the trajectories of*

$$D_s\dot{x} = \frac{1}{2}\,\mathrm{grad}_x H(x, s)$$

are also complete.

Recall that the completeness of M is an obvious necessary condition (the wave fronts are totally geodesic) and, then, the question is fully reduced to a purely Riemanian problem: the completeness of the trajectories of the potential $V = -H/2$. This problem was studied by several authors in the 1970s [18, 25, 46] and they obtained very accurate results when the potential is autonomous, i.e., H independent of u. For example, a result by Weinstein and Marsden [46] (see also [1, Theorem 3.7.15] or [15, Sect. 3]), formulated in terms of positively complete functions, yields as a straightforward consequence:

Theorem 3. *Any PFW with M complete and $H(x, u) \equiv H(x)$ at most quadratic is geodesically complete.*

Recall that here only H (and not $-H$) needs to be controlled. As an example of the power of this result, one can check that the explicit examples of pp-waves exhibited in [29], which were proven to be complete (by integrating

[4] The subscript Δ for V (and similarly in formula (5) below, for $\mathcal{J}$) is only an abbreviation to keep track of Δu. Notice that the dependence of V on Δu will be important for geodesic connectedness and the existence of conjugate points, but we will get rid of it for geodesic completeness.

– decoupled – geodesic equations), lie under the hypotheses[5] of Theorem 3. For example, for PP1 (see [29, Sect. 5.2]), $H(x) = \cos x^2 - \cosh x^1$; for PP2, $H(x) = -\sum_j f_j(x^j)$ with the f_j's bounded from below; in both cases, H is upper bounded. On the other hand, their incomplete examples strongly violate the conditions of Theorem 3. For example, for the monopole pp-waves in PP3 the Riemannian part M may be incomplete, and for the example PP4 one has the highly violating coefficient $H(x) = -e^{x^2} \sin x^1$.

Nevertheless, the results for non-autonomous potentials are not so accurate [25]. But this is the case of plane waves, which are geodesically complete in any dimension (see [15, Proposition 3.5]) and, then, to find general and accurate criteria seems an interesting topic for further research.

4.2 Ehlers–Kundt Question

From a fundamental viewpoint, the following question on pp-waves ($M = \mathbb{R}^2$) was posed by Ehlers and Kundt [19] (see also [8] or [29]):

> Is any *complete* gravitational pp-wave a plane wave?

As they pointed out, complete gravitational pp-waves represent graviton fields generated independently of matter (vacuum) or external sources (completeness). Thus, they are analogous to source-free photons in electrodynamics.

Notice that the hypotheses become relevant for both the physical interpretation and the involved mathematical problem. In fact, from Theorem 2 (with $V = -H/2$) and the fact that linear terms in the expression of H in (2) can be dropped by choosing appropriate coordinates, the previous question is equivalent to:

> Let $V((x, y), s)$, $V : \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}$ be an harmonic function in (x, y). If the trajectories for V as a (non-autonomous) potential on $\mathbb{R}^2$ are complete, must V be a (harmonic) polynomial of degree ≤ 2 for each fixed s (i.e., $V((x, y), s) = f(s)(x^2 - y^2) + 2g(s)xy + c(s)x + d(s)y + e(s)$)?

Notice that the harmonicity of H allows us to use techniques from complex analysis. In fact, it is easy to show:

Theorem 4. *Any gravitational pp-wave such that $H(x, u)$ behaves at most quadratically at spatial infinity is a (necessarily complete) plane wave.*

To prove this, put $\zeta = x + iy$, $H \equiv H(\zeta, u)$ and consider the complex function $f(\zeta, u)$ which is holomorphic in ζ with real part equal to H. Then, $f(\zeta, u)/\zeta^2$ is meromorphic for $\zeta \in \mathbb{C}$ and bounded for big ζ. Thus, for each u, whenever $f(\cdot, u)$ is not constant, it presents a pole at infinity of order 1 or 2. That is,

[5] For the comparison of hypotheses, recall that their function $F(x, u)$ plays the role of our $-H(x, u)$.

$f(\cdot, u)$ is a complex polynomial of degree at most 2, and the result follows directly.

Even though Theorem 4 covers the most meaningful cases from the physical viewpoint (and is free of hypotheses on completeness), in general the above question remains open as a mathematical problem with roots in the foundations of the theory of gravitational waves.

5 Geodesic Connectedness and Conjugate Points

5.1 The Lorentzian Problem

Next, we will study geodesic connectedness of PFWs, that is, we will ask: fixed any $z_0 = (x_0, v_0, u_0), z_1 = (x_1, v_1, u_1) \in \mathcal{M}$, is there any geodesic connecting z_0, z_1? This problem becomes relevant from different viewpoints (see [41] for a survey): (a) the connectivity of a point z_0 with any point $z_1 \in I^+(z_0)$ through a timelike geodesic admits an obvious physical interpretation, and is satisfied by all globally hyperbolic spacetimes (Avez–Seifert result), (b) the geodesic connectedness of a Lorentzian manifold – through geodesics of any causal type – is a desirable geometrical property[6], which admits a natural variational interpretation and, then, yields an excellent motivation to study critical points of indefinite functionals from a mathematical viewpoint [34], (c) the possible multiplicity of connecting geodesics is related to the existence of conjugate points.

These questions were studied by Penrose [37] and Ehrlich and Emch [20, 21, 22] for plane waves, by integrating geodesic equations. They proved that there exists a natural concept of *conjugacy for pairs* $u_0, u_1 \in \mathbb{R}$, $u_0 < u_1$, and obtained the following results:

1. (Penrose). Lightlike geodesics are focused when u_0, u_1 are conjugate (at least for "weak" sandwich waves). In this case, all the lightlike geodesics starting at z_0 (except one)
   – either cross a fixed point with $u = u_1$ (anastigmatic conjugacy, in electromagnetic plane waves),
   – or cross a fixed line (astigmatic conjugacy, in gravitational or mixed plane waves).
2. (Ehrlich-Emch). The connectable points for astigmatic gravitational plane waves can be characterized in an accurate way:
   – if u_1 lies before the first conjugate point of u_0, then there exists a unique geodesic between z_0 and z_1, which is causal if $z_0 < z_1$;
   – otherwise, connecting geodesics may not exist and, in fact, *gravitational plane waves are not geodesically connected.*

[6] This is trivially satisfied for complete Riemannian manifolds but not necessarily for complete Lorentzian ones, such as de Sitter spacetime.

5.2 Relation with a Purely Riemannian Variational Problem

From the study of geodesic equations in Sect. 4, and the classical relation between connecting trajectories for a potential and extremal of Lagrangians, it is not difficult to prove [15]:

Proposition 3. *For fixed $z_0, z_1 \in \mathcal{M}$, the following statements are equivalent:*

(a) *z_0 and z_1 can be connected by a geodesic.*
(b) *There exists a solution for the Riemannian problem*

$$\begin{cases} D_s\dot{x}(s) = -\mathrm{grad}_x V_\Delta(x(s), s) & \text{for all } s \in [0,1] \\ x(0) = x_0, \;\; x(1) = x_1 \,, \end{cases}$$

where $V_\Delta(x,s) = -\frac{(\Delta u)^2}{2}\, H(x, u_0 + s\Delta u)$, $\Delta u = u_1 - u_0$.
(c) *There exists a critical point for the action functional $\mathcal{J}_\Delta$ defined on the space of absolutely continuous curves $x : [0,1] \to M$ which connect x_0, x_1,*

$$\mathcal{J}_\Delta(x) = \frac{1}{2} \int_0^1 \langle \dot{x}, \dot{x} \rangle_R \; ds \; - \int_0^1 V_\Delta(x,s) \; ds \,. \tag{5}$$

Of course, (c) is *the most classical problem in Lagrangian Mechanics*. Nevertheless (as a surprise for us), it had not been fully solved in the quadratic case. This case corresponds to plane waves and, thus, in order to obtain optimal Lorentzian results (re-obtaining in particular Ehrlich-Emch's), we had to improve the known Riemannian ones. The final Riemannian result [13] is the following (see the Appendix for a discussion of the problem):

Theorem 5. *Let $(M, \langle\cdot,\cdot\rangle_R)$ be a complete (connected) n-dimensional Riemannian manifold. Assume that $V \in C^1(M \times [0,1], \mathbb{R})$ is at most quadratic in x in the following way:*

$$V(x,s) \leq \lambda d^2(x, \bar{x}) + \mu d^p(x, \bar{x}) + k \quad \forall (x,s) \in M \times [0,1] \,,$$

for some fixed point $\bar{x} \in M$ and constants $p < 2$, $\lambda, \mu, k \geq 0$.

If $\lambda < \pi^2/2$ then, for all $x_0, x_1 \in M$, there exists at least one critical point (in fact, an absolute minimum) of $\mathcal{J}_\Delta$ in (5). In particular, this happens if V is subquadratic, i.e., when $\lambda = 0$.

If, additionally, M is not contractible, then there exists a sequence of critical points $\{x_m\}_m$ such that

$$\lim_{m \to +\infty} \mathcal{J}_\Delta(x_m) \to +\infty \,.$$

One can also assume that λ (as well as p, μ, k) depend on s, and then take the maximum of $\lambda([0,1])$ (and the others) for the conclusion.

5.3 Optimal Results for Connectedness of PFWs

Now, the application of Proposition 3 and Theorem 5 (plus a further discussion for the case of causal geodesics) directly yields the following result. Notice that the limit value $\lambda = \pi^2/2$ in Theorem 5 is related to $(\Delta u)^2$ in the expression of V_Δ in Proposition 3(b); thus, in the at most quadratic case, stronger conclusions hold when $(\Delta u)^2$ is smaller than a critical value.

Theorem 6. *Let $\mathcal{M}$ be a PFW with M complete, and fix $\bar{x} \in M$. Then:*

(1) *If $-H(x,u)$ is spatially subquadratic, then $\mathcal{M}$ is geodesically connected.*
(2) *If $-H(x,u)$ is at most quadratic with*

$$-H(x,u) \leq R_0(u)\, d^2(x,\bar{x}) + R_1(u)\, d^{p(u)}(x,\bar{x}) + R_2(u)$$

$\forall (x,u) \in M \times \mathbb{R}$, $p(u) < 2$, then $z_0 = (x_0, v_0, u_0)$, $z_1 = (x_1, v_1, u_1) \in \mathcal{M}$, $u_0 \leq u_1$ can be connected by means of a geodesic whenever

$$R_0[u_0,u_1]\,(u_1-u_0)^2 < \pi^2\ ,$$

where

$$R_0[u_0,u_1] = \mathrm{Max}\{R_0(u) : u \in [u_0,u_1]\}\ .$$

Moreover, in any of the previous cases (1), (2)*:*

(a) *If $z_0 < z_1$ there exists a length-maximizing causal geodesic connecting z_0 and z_1;*
(b) *If M is not contractible:*
  (i) *There exist infinitely many spacelike geodesics connecting z_0 and z_1,*
  (ii) *The number of timelike geodesics from z_0 to $z_v = (x_1, v, u_1)$ goes to infinity when $-v \to \infty$.*

It must be emphasized that these results are optimal because the Riemannian results are optimal. In fact:

- There are explicit counterexamples if any of the hypotheses is dropped.
- In the case of gravitational plane waves, the conclusions of Theorem 6 not only generalize Ehrlich-Emch's ones, but also yield bounds for the appearance of the first astigmatic conjugate pair – a lower bound is the value u_+ ($u_+ > u_0$) such that $R_0[u_0,u_+](u_+-u_0)^2 = \pi^2$.
- All the results can be naturally extended to the case of M being non-complete with convex boundary.

5.4 Conjugate Points

From the above approach to geodesic connectedness it is also clear that, now, the existence of conjugate points for geodesics on a PFW is equivalent to the existence of conjugate points for the action $\mathcal{J}_\Delta$. More precisely, following [24, Sect. 6], we can define:

Definition 1. *Fix* $\overline{z}_0 = (x_0, u_0)$, $\overline{z}_1 = (x_1, u_1) \in M \times \mathbb{R}$, *and let* $x(s)$ *be a critical point of* $\mathcal{J}_\Delta$ *in (5) with endpoints* x_0, x_1 *and* $\Delta u = u_1 - u_0$. *We say that* $\overline{z}_0$, $\overline{z}_1$ *are conjugate points along* $x(s)$ *of multiplicity* $\overline{m}$ *if the dimension of the nullity of the Hessian of* $\mathcal{J}_\Delta$ *on* $x(s)$ *is* $\overline{m}$ (*if* $\overline{m} = 0$ *we say that* $\overline{z}_0$, $\overline{z}_1$ *are not conjugate*).

Then, one obtains the following equivalence between conjugate points for Lorentzian geodesics and conjugate points for Riemannian trajectories of a potential [24, Proposition 6.2]:

Proposition 4. *The pairs* $\overline{z}_0 = (x_0, u_0)$, $\overline{z}_1 = (x_1, u_1)$ *are conjugate of multiplicity* $\overline{m}$ *along* $x(s)$ *if and only if for any geodesic* $z : [0, 1] \to \mathcal{M}$ *with* $z(s) = (x(s), v(s), \Delta u \cdot s + u_0)$ *the corresponding endpoints* $z_0 = (x_0, v_0, u_0)$, $z_1 = (x_1, v_1, u_1)$ *are conjugate with the same multiplicity* $m = \overline{m}$.

As we mentioned in Subsect. 5.1, in the particular case of gravitational plane waves, conjugate pairs are defined for u_0, u_1. For general PFWs, the lack of symmetries of the fronts makes it necessary to take care of the M part. Nevertheless, the dependence on v is still dropped.

Now, studying the conjugate points for $\mathcal{J}_\Delta$, one can obtain easily results as [24, Proposition 6.4]: *If* H *is spatially convex* (i.e. $\mathrm{Hess}_x H(x, u)(w, w) \geq 0$, $\forall w \in T_x M$) *and the sectional curvature of* M *is non-positive, then no geodesic admits conjugate points.* Of course, the hypotheses of this result go in the *wrong* direction with respect to the energy conditions ($\Delta_x H \leq 0, K \geq 0$), which tend to yield conjugate points. Nevertheless, this focusing of geodesics is, in general, qualitatively different from the focusing in the plane-wave case and, as we have seen, it does not forbid global hyperbolicity.

Appendix: The Riemannian Problem of Connectedness by the Trajectories of a Lagrangian

In Sect. 5 we showed that geodesic connectedness of PFWs depends crucially on the Riemannian variational result Theorem 5. This result is an answer to the classical Bolza problem, which can be stated as:

> **Bolza problem.** For fixed x_0, x_1 in a Riemannian manifold M and some $T > 0$, determine the existence of critical points for the functional
>
> $$J_T(x) = \frac{1}{2} \int_0^T \langle \dot{x}, \dot{x} \rangle_R \, ds - \int_0^T V(x, s) \, ds$$
>
> on the set of absolutely continuous curves with $x(0) = x_0, x(T) = x_1$.

In our case, $T = 1$, V is smooth and at most quadratic, and M is complete. For this problem, it is well-known that two abstract conditions on J_T, namely, boundedness from below and coercivity, imply the existence of a critical point

– in fact a minimum. Even more, by using Ljusternik–Schnirelmann theory one can ensure the existence of a sequence of critical points such that J_T diverges.

The following results were known:

1. If $V(x,s)$ is bounded from above or subquadratic in x, then the two abstract conditions hold and J_T attains a minimum.
2. If $V(x,s)$ is at most quadratic, with $V(x,s) \leq \lambda d^2(x,\bar{x}) + \mu d^{p(s)}(x,\bar{x}) + k(s),\ p(s) < 2, \forall s \in [0,T]$ then:
   - Clarke and Ekeland [16] proved that, if $T < 1/\sqrt{\lambda}$, then J_T still admits a minimum.
   - If $T \geq \pi/\sqrt{2\lambda}$ there are simple counterexamples to the existence of critical points (harmonic oscillator).

Therefore, there was a gap for the values of T,

$$T \in [\,1/\sqrt{\lambda}, \pi/\sqrt{2\lambda}\,[\,,$$

which was covered only in some particular cases (for example, if $\mathrm{Hess}_x V \geq 2\lambda$, then J_T still admits a minimum). Our results in [13] (essentially, Theorem 5) fill this gap, by showing that, even in the case $T \in [\,1/\sqrt{\lambda}, \pi/\sqrt{2\lambda}\,[$, the functional J_T is bounded from below and coercive and, thus, admits a minimum.

The proof was carried out in three steps:

– Step 1. The essential term to prove the abstract conditions for J_T is $d^2(x(s),\bar{x})$. Then, consider the new functional

$$F_T^\lambda(x) = \frac{1}{2}\int_0^T \langle \dot{x}, \dot{x}\rangle_R\, ds \;-\; \lambda \int_0^T d^2(x(s),\bar{x})\, ds\,.$$

J_T is essentially greater than F_T^λ and, if F_T^λ is bounded from below and coercive, then so is J_T (recall that the expression of F_T^λ contains $d^2(\cdot,\bar{x})$, which is only continuous, but we are not looking for critical points of this functional).

– Step 2. Reduction to a problem in one variable. For each curve $x(s)$ in the domain of J_T, one can find a continuous curve $y(s), s \in [0,T]$, almost everywhere differentiable, such that (assuming $\bar{x} = x_0$ without loss of generality) $y(0) = 0$, $y(T) = d(x_0,x_1)$ and:

$$\dot{y}(s) = |\dot{x}(s)| \text{ a.e. in } [0,s_0], \qquad \dot{y}(s) = -|\dot{x}(s)| \text{ a.e. in }]\,s_0,T\,]\,,$$

for some suitable s_0. For this curve $y(s)$,

$$F_T^\lambda(x) \geq \frac{1}{2}\int_0^T |\dot{y}|^2\, ds \;-\; \lambda \int_0^T |y|^2\, ds\,. \tag{6}$$

And, then, one has just to prove that the new (1-*dimensional*) functional $G_T^\lambda(y)$, equal to the right hand side of (6), is coercive and bounded from below.

– Step 3. Solution of the 1-variable problem for $G_T^\lambda(y)$ by elementary methods (Fourier series, Wirtinger's inequality).

The technique also works for manifolds with boundary [14]. Remarkably, the procedure has also been used to prove the geodesic connectedness of static spacetimes under critical quadratic hypotheses [3] (see also [42]), and other related problems.

Acknowledgements

J.L.F. has been supported by a MECyD Grant EX-2002-0612. M.S. has been partially supported by a MCyT-FEDER Grant MTM2004-04934-C04-01.

References

1. R. Abraham, J. Marsden: *Foundations of Mechanics*, 2nd edn. (Addison–Wesley, Massachusetts 1978) 90
2. D. Amati, C. Klimcik: Phys. Lett. B **219**, 443 (1989) 80
3. R. Bartolo, A.M. Candela, J.L. Flores, M. Sánchez: Adv. Nonlin. Stud. **3**, 471 (2003) 97
4. J.K. Beem, P.E. Ehrlich, K.L. Easley: *Global Lorentzian Geometry* (Marcel Dekker, New York 1996) 81, 83, 84, 87
5. D. Berenstein, J. Maldacena, H. Nastase: J. High Energy Physics **0204**, 013 (2002). `hep-th/0202021` 80
6. A.N. Bernal, M. Sánchez: Commun. Math. Phys. **243**, 461 (2003) 87
7. A.N. Bernal, M. Sánchez: Smoothness of time functions an the metric splitting of globally hyperbolic spacetimes, Commun. Math. Phys. (to appear). `gr-qc/0401112` 87
8. J. Bičak: Selected solutions of Einstein's field equations: Their role in general relativity and astrophysics. In: *Einstein's Field Equations and Their Physical Interpretations*, Lect. Notes Phys. **540**, ed by B. Schmidt (Springer, Heidelberg 2000) pp 1–126. `gr-qc/0004016` 83, 91
9. M. Blau, M. Borunda et al: Class. Quant. Grav. **21**, L43 (2004). `hep-th/0312029` 80, 85
10. M. Blau, J. Figueroa-O'Farrill, G. Papadopoulos: Class. Quant. Grav. **19**, 4753 (2002). `hep-th/0202111` 80
11. E. Boyda, S. Ganguli et al: Phys. Rev. D **67**, 106003 (2003). `hep-th/0212087` 80
12. H. Brinkmann: Math. Ann. **94**, 119 (1925) 80
13. A.M. Candela, J. L. Flores, M. Sánchez: J. Diff. Equat. **193**, 196 (2003) 82, 93, 96
14. A.M. Candela, J.L. Flores, M. Sánchez: Discr. and Contin. Dyn. Syst., added vol., 173 (2003) 97
15. A.M. Candela, J.L. Flores, M. Sánchez: Gen. Rel. Grav. **35**, 631 (2003). `gr-qc/0211017` 80, 81, 82, 84, 89, 90, 91, 93
16. H.F. Clarke, I. Ekeland: Arch. Rational Mech. Anal. **78**, 315 (1982) 96
17. A. Coley, R. Milson et al: Phys. Rev. D **67**, 104020 (2003). `gr-qc/0212063` 80

18. D. Ebin: Proc. Amer. Math. Soc. **26**, 632 (1970) 90
19. J. Ehlers, K. Kundt: Exact solutions of the gravitational field equations. In: *Gravitation: An Introduction to Current Research*, ed by L. Witten (J. Wiley & Sons, New York 1962) pp 49–101 80, 82, 83, 91
20. P.E. Ehrlich, G.G. Emch: Rev. Math. Phys. **4**, 163 (1992) 81, 92
21. P.E. Ehrlich, G.G. Emch: Lecture Notes in Pure and Appl. Math. **144**, 203 (1992) 81, 92
22. P.E. Ehrlich, G.G. Emch: Proc. Symp. Pure Math. **54**, 203 (1993) 81, 92
23. A. Einstein, N. Rosen: J. Franklin Inst. **223**, 43 (1937) 80
24. J.L. Flores, M. Sánchez: Class. Quant. Grav. **20**, 2275 (2003). `gr-qc/0211086` 81, 82, 84, 85, 86, 87, 88, 94, 95
25. W.B. Gordon: Proc. Amer. Math. Soc. **26**, 329 (1970) 90, 91
26. T. Harmark, T. Takayanagi: Nucl. Phys. B **662**, 3 (2003). `hep-th/0301206` 80
27. S.W. Hawking, G.F.R. Ellis: *The Large Scale Structure of Space-Time*, (Cambridge University Press, Cambridge 1973) 83, 87
28. G.T. Horowitz, A.R. Steif: Phys. Rev. Lett. **64**, 260 (1990) 80
29. V.E. Hubeny, M. Rangamani: J. High Energy Physics **0212**, 043 (2002). `hep-th/0211195` 82, 90, 91
30. V.E. Hubeny, M. Rangamani: J. High Energy Physics **0211**, 021 (2002). `hep-th/0210234` 82, 85, 89
31. V.E. Hubeny, M. Rangamani: Mod. Phys. Lett. **A18**, 2699 (2003) 82, 85, 89
32. V.E. Hubeny, M. Rangamani, S. F. Ross: Phys. Rev. D **69**, 024007 (2004). `hep-th/0307257` 82, 87, 88
33. L. Maoz, J. Simón: J. High Energy Physics **0401**, 051 (2004). `hep-th/0310255` 80, 82, 87
34. A. Masiello: *Variational Methods in Lorentzian Geometry*, Pitman Res. Notes Math. Ser. **309** (Longman Sci. Tech., Harlow 1994) 92
35. C.W. Misner, K.S. Thorne, J.A. Wheeler: *Gravitation*, (Freeman, San Francisco 1973) 80
36. B. O'Neill: *Semi-Riemannian Geometry with Applications to Relativity*, (Academic Press, New York 1983) 83
37. R. Penrose: Rev. Mod. Phys. **37**, 215 (1965) 80, 81, 82, 85, 92
38. R. Penrose: Singularities and time-asymmetry. In: *General Relativity, an Einstein Centenary Survey*, ed by S.W. Hawking, W. Israel (Cambridge Univ. Press, Cambridge 1979) pp 581–638 81
39. R. Penrose: Any spacetime has a plane wave as a limit. In: *Differential Geometry and Relativity*, ed by M. Cahen, M. Flato (Reidel, Dordrecht 1976) pp 271–275 80
40. V. Pravda, A. Pravdova et al: Class. Quant. Grav. **19**, 6213 (2002). `gr-qc/0209024` 80
41. M. Sánchez: Nonlinear Anal. **47**, 3085 (2001) 92
42. M. Sánchez: On the geometry of static spacetimes, Nonlinear Anal. (to appear). `dg/0406332` 97
43. M. Sánchez: Causal hierarchy of spacetimes, temporal functions and smoothness of Geroch's splitting. A revision. In: *Proceedings of the* 13*th School of Differential Geometry*, São Paulo, Brazil, 2004 (to appear in *Matemática Contemporânea*). `gr-qc/0411143` 87
44. J.M.M. Senovilla: J. High Energy Physics **0311**, 046 (2003). `hep-th/0311172` 82, 89
45. H. Stephani, D. Kramer, M. MacCallum, C. Hoenselaers, E. Herlt: *Exact Solutions of Einstein's Field Equations* (Cambridge University Press, Cambrigde 2003) 83
46. A. Weinstein, J. Marsden: Proc. Amer. Math. Soc. **26**, 629 (1970) 90

Part II

Analytical Methods and Differential Equations

Concepts of Hyperbolicity and Relativistic Continuum Mechanics

Robert Beig

Institut für Theoretische Physik der Universität Wien, Boltzmanngasse 5, 1090 Wien, Austria
robert.beig@univie.ac.at

Abstract. After a short introduction to the characteristic geometry underlying weakly hyperbolic systems of partial differential equations we review the notion of symmetric hyperbolicity of first-order systems and that of regular hyperbolicity of second-order systems. Numerous examples are provided, mainly taken from nonrelativistic and relativistic continuum mechanics.

1 Introduction

The notion of hyperbolicity of a partial differential equation (PDE), or a system of PDE's, is central for the field theories of mathematical physics. It is closely related to the well-posedness of the Cauchy problem and to the causal structure underlying these theories. In standard theories describing relativistic fields in vacuo this causal structure is that given by the spacetime metric, a second-order symmetric tensor of Lorentzian signature. If matter is included, things become both more complicated and more subtle. In fact, the awareness of some of those complications predates Relativity by centuries. An example is afforded by the phenomenon, already studied by Huygens, of birefringence in crystal optics[1].

There is currently an increase of attention in the field of Relativity, due in part to demands from Numerical Relativity, devoted to certain notions of hyperbolicity applied to the Einstein equations (for an excellent review see [18]). There the main challenge, not discussed in the present notes at all, comes from the fact that, already in vacuum, the Einstein equations by themselves, i.e. prior to the imposition of any gauge conditions, are not hyperbolic. The main burden, then, is to find a "hyperbolic reduction" turning the Einstein equations, or a subset thereof, into a hyperbolic system appropriate for the purpose at hand. However the complications in the causal structure one finds in continuum mechanics, which are our main focus here, are absent in the Einstein vacuum case – at least for the reductions proposed so far. Of course, these complications do come into play ultimately once matter-couplings are included.

[1] For a fascinating account of the history of the associated mathematics see [20].

R. Beig: *Concepts of Hyperbolicity and Relativistic Continuum Mechanics*, Lect. Notes Phys. **692**, 101–116 (2006)
www.springerlink.com

These notes attempt an elementary introduction to some notions of hyperbolicity and the "characteristic geometry" associated with or underlying these notions. The section following this one is devoted to the general notion of a hyperbolic polynomial, which in our case of course arises as the characteristic polynomial of a PDE. It is interesting that this notion is on one hand restrictive enough to encode essentially all the required features of a theory in order to be "causal" – on the other hand flexible enough to account for an amazing variety of phenomena – relativistic or nonrelativistic – ranging from gravitational radiation to water waves or phonons in a crystal. We devote a significant fraction of Sect. 2 to examples, which at least in their nonrelativistic guise all appear in the standard literature such as [13], though not perhaps from the unified viewpoint pursued here. Some of these examples are not fully worked out, but perhaps the interested reader is encouraged to fill in more details, possibly using some of the cited literature. We hope that some workers in Relativity, even if they have little interest in continuum mechanics for its own sake, find these examples useful for their understanding of the notion of hyperbolicity. While hyperbolicity of the characteristic polynomial of a theory is important, it is not in general sufficient for the well-posedness of the initial value problem for that theory. Well-posedness is the subject of our Sect. 3. We recall the notion of a symmetric hyperbolic system of a system of 1st order PDE's, which is indeed sufficient for well-posedness. A similar role for 2nd order equations is played by a class of systems, which were to some extent implicit in the literature, and for which an elaborate theory has been recently developed in [10, 11]. These systems are called regular hyperbolic. They encompass many second order systems arising in physics one would like to qualify as being hyperbolic – such as the Einstein equations in the harmonic gauge. If applicable, the notion of regular hyperbolicity is particularly natural for systems of 2nd order derivable from an action principle, as is the case for many problems of continuum mechanics. We show the fact, obvious for symmetric hyperbolic systems and easy-to-see although not completely trivial for regular hyperbolic ones, that these systems are special cases of weakly hyperbolic systems, i.e. ones the determinant of whose principal symbol is a hyperbolic polynomial. We also touch the question of whether a system of the latter type can be reduced to one of the former type by increasing the number of dependent variables. Throughout this section our treatment will be informal in the sense of ignoring specific differentiability requirements. We also do not touch questions of global well-posedness.

2 Hyperbolic Polynomials

The PDE's we are interested in are of the form

$$M_{AB}^{\mu_1\cdots\mu_l}(x, f, \partial f, \ldots, \partial^{(l-1)} f)\partial_{\mu_1}\ldots\partial_{\mu_l} f^B + \text{lower order terms} = 0\,. \quad (1)$$

Here $A, B = 1, ..., m$ and $\mu_i = 1, \ldots, n$. Relevant equations of this form are the Euler equations for a barotropic fluid (for $n = 4, l = 1, m = 4$), the Einstein equations (for $n = 4, l = 2, m = 10$) or the equations governing an ideal elastic solid (for $n = 4, l = 2, m = 3$). The Maxwell equations, in the form they are originally written down, are not of this form, but a suitable subset of them is, as we will discuss later.

The principal symbol of the PDE (1) is defined as

$$M_{AB}(k) = M_{AB}^{\mu_1 \cdots \mu_l} k_{\mu_1} \ldots k_{\mu_l}, \quad k_\mu \in (\mathbb{R}^n)^* \tag{2}$$

We here suppress the dependence on x and on f. The characteristic polynomial $P(k)$ is defined by $P(k) = det\, M_{AB}(k)$, where the determinant is taken with respect to some volume form on f-space $\subset \mathbb{R}^m$. The polynomial $P(k)$ is homogenous of degree $p = m \cdot l$. A homogenous polynomial of degree $p > 0$ is called hyperbolic with respect to $\xi_\mu \in (\mathbb{R}^n)^*$ if $P(\xi) \neq 0$ and the map $\lambda \mapsto P(\eta + \lambda\xi)$, itself a polynomial of degree p, has only real roots $\lambda_i, i = 1, \cdots, p$ for all $\eta \in (\mathbb{R}^n)^*$. The roots $\lambda_i(\xi, \eta)$ need not be distinct. If, for all η with $\eta \wedge \xi \neq 0$, $\lambda_i(\xi, \eta) \neq \lambda_j(\xi, \eta)$ for $i \neq j$, P is called strictly hyperbolic[2]. We write $\mathcal{C}^*$ for the set of $k \in (\mathbb{R}^n)^* \setminus \{0\}$, where P vanishes. It is sometimes called the cone of characteristic conormals.

It is clear that a product of hyperbolic polynomials is hyperbolic. Also, if a hyperbolic polynomial can be factorized into polynomials of lower degree (in which case it is called reducible), these factors are also hyperbolic. There is a wealth of information which can be inferred about a polynomial $P(k)$ if it is hyperbolic. Before explaining some of this, we look at a few examples for hyperbolic polynomials.

Example 1. $P(k) = (a, k) = a^\mu k_\mu$ for some nonzero $a^\mu \in \mathbb{R}^n$. The set $\mathcal{C}^*$ is a punctured hyperplane $\subset (\mathbb{R}^n)^*$.

Clearly $P(k)$ is hyperbolic with respect to any ξ_μ such that $a^\mu \xi_\mu \neq 0$. The polynomial $P(k) = (a_1, k)(a_2, k)(a_3, k)$, with a_1, a_2, a_3 linearly independent $\in \mathbb{R}^3$, arises in the problem of finding, for a three dimensional positive definite metric, a coordinate system in which the metric is diagonal (see [14]) – which shows that hyperbolic problems can also arise in purely Riemannian contexts.

Example 2. $P(k) = \gamma^{\mu\nu} k_\mu k_\nu$, where $\gamma^{\mu\nu}$ is a (contravariant) metric of Lorentzian signature $(-, +, \ldots, +)$. The set $\mathcal{C}^*$ is the two-sheeted Minkowski light cone.

When $n = 2$, $P(k)$ is hyperbolic with respect to any non-null ξ_μ, when $n > 2$, $P(k)$ is hyperbolic with respect to any ξ_μ with $\gamma^{\mu\nu} \xi_\mu \xi_\nu < 0$, i.e. ξ is timelike with respect to $\gamma^{\mu\nu}$. Checking that $P(k)$ is hyperbolic according to

[2] This case is not general enough for the purposes of physics. Furthermore there exist physically relevant cases of non-strictly hyperbolic polynomials which are stable, in the sense that they possess open neighbourhoods in the set of hyperbolic polynomials just containing non-strictly hyperbolic ones [28, 26].

our definition is equivalent to the so-called reverse Cauchy-Schwarz inequality for two covectors one of which is timelike or null with respect to $\gamma^{\mu\nu}$ (which is the mathematical rationale behind the twin "paradox" of Relativity). Surprisingly there are similar inequalities for general hyperbolic polynomials (see [19]) which play a role in diverse fields of mathematics [5].

Example 2 is of course the most familiar one. If it arises from nonrelativistic field theory, the quantity $\gamma^{\mu\nu}$ currently runs under the name of the "Unruh or acoustic metric" [4] (see also [12]) in the Relativity community. It is not an elementary object of the theory, but is built as follows: Take first the Galilean metric $h^{\mu\nu}$, a symmetric tensor with signature $(0,+,\ldots,+)$ together with a nonzero covector τ_μ satisfying $h^{\mu\nu}\tau_\nu = 0$: these are the absolute elements . Then pick a 4-vector u^μ normalized so that $u^\mu\tau_\mu = 1$ and define $\gamma^{\mu\nu} = h^{\mu\nu} - c^{-2}u^\mu u^\nu$. This describes waves propagating isotropically at phase velocity c in the rest system, defined by u^μ, of a material medium. The relativistic version of the above is as follows: Start with the spacetime metric $g^{\mu\nu}$ and define $\gamma^{\mu\nu} = g^{\mu\nu} + (1-c^{-2})u^\mu u^\nu$, where u^μ is normalized by taking $\tau_\mu = -g_{\mu\nu}u^\nu$, with $g_{\mu\nu}$ the covariant spacetime metric defined by $g_{\mu\nu}g^{\nu\lambda} = \delta_\mu{}^\lambda$. Note: if there are metrics $\gamma_1^{\mu\nu}$,$\gamma_2^{\mu\nu}$ with $c_2 < c_1$, then the "faster" cone lies inside the slower one. We will come back to this point later.

Example 3. $P(k) = s^{\mu\nu}k_\mu k_\nu$, where $s^{\mu\nu}$ has signature $(-,+,...,+,0,...,0)$, is hyperbolic with respect to any ξ such that $s^{\mu\nu}\xi_\mu\xi_\nu < 0$.

Here is a case occurring in the real world. Let $g_{\mu\nu}$ be a Lorentz metric on $\mathbb{R}^4$, u^μ a normalized timelike vector, i.e. $g_{\mu\nu}u^\mu u^\nu = -1$, $F_{\mu\nu} = F_{[\mu\nu]}$ nonzero with $F_{\mu\nu}u^\nu = 0$. The quadratic form $s^{\mu\nu} = -eu^\mu u^\nu + 1/2\ F_{\rho\sigma}F^{\rho\sigma}g^{\mu\nu} - F^\mu{}_\rho F^{\nu\rho}$, with $e > 0$, has signature $(-,+,+,0)$. The characteristic cone $\mathcal{C}^*$ of $P(k) = 0$ consists of two hyperplanes punctured at the origin. When e is interpreted as $e =$"energy density + pressure" and $F_{\mu\nu}$ as the frozen-in magnetic field of an ideally conducting plasma, then $P(k)$ describes the Alfvén modes of relativistic magnetohydrodynamics [45][2].

Example 4. Let $n = 4$, $\varepsilon_{\mu\nu\lambda\rho}$ some volume form on $\mathbb{R}^4$ and $m^{\mu\nu\lambda\rho} = m^{[\mu\nu][\lambda\rho]}$. With $G^{\mu\nu\rho\sigma} = \varepsilon_{\alpha\beta\delta\epsilon}\,\varepsilon_{\kappa\phi\psi\omega}\,m^{\alpha\beta\kappa(\mu}m^{\nu|\delta\phi|\rho}m^{\sigma)\epsilon\psi\omega}$ we define P by $P(k) = G^{\mu\nu\rho\sigma}k_\mu k_\nu k_\rho k_\sigma$.

As a special case take $m^{\mu\nu\lambda\rho}$ of the form $m^{\mu\nu\lambda\rho} = h^{\lambda[\mu}h^{\nu]\rho} - e^{\lambda[\mu}u^{\nu]}u^\rho + e^{\rho[\mu}u^{\nu]}u^\lambda$, where the symmetric tensors $h^{\mu\nu}$ and $s^{\mu\nu}$, both of signature $(0,++ +)$, satisfy $h^{\mu\nu}\tau_\nu = e^{\mu\nu}\tau_\nu = 0$ for $u^\mu\tau_\mu \neq 0$: this is the situation encountered in crystal optics with the nonzero eigenvalues of $e^{\mu\nu}$ relative to $h^{\mu\nu}$ being essentially the dielectric constants. The crystal is optically biaxial or triaxial, depending on the number of mutually different eigenvalues. The 4th order polynomial $P(k)$ turns out to be hyperbolic with respect to all ξ_μ in some neighbourhood of $\xi_\mu = \tau_\mu$, and the associated characteristic cone is the Fresnel surface (see e.g. [25]). For an optically isotropic medium or in vacuo $P(k)$ is reducible, in fact the square of a quadratic polynomial of the type of

Example 2. We leave the details as an exercise. More general conditions on $m^{\mu\nu\lambda\rho}$ in order for $P(k)$ to be hyperbolic can be inferred from [32].

The quartic polynomial $P(k)$, as defined above, comes from a generalized ("pre-metric") version of electrodynamics (see [24]), as follows: Let $F_{\mu\nu}$ be the electromagnetic field strength and write $H^{\mu\nu} = m^{\mu\nu\lambda\rho}F_{\lambda\rho}$ for the electromagnetic excitation. The premetric Maxwell equations then take the form

$$\partial_{[\mu}(\varepsilon_{\nu\lambda]\rho\sigma}H^{\rho\sigma}) = J_{\mu\nu\lambda}, \qquad \partial_{[\mu}F_{\nu\lambda]} = 0, \tag{3}$$

where $J_{\mu\nu\lambda}$ is the charge three form[3]. The (3) reduce to the standard ones in vacuo when $m^{\mu\nu\lambda\rho} \sim g^{\lambda[\mu}g^{\nu]\rho}$ with $g^{\mu\nu}$ the metric of spacetime. If one sets $m^{\mu\nu\lambda\rho} = \gamma^{\lambda[\mu}\gamma^{\nu]\rho}$, with $\gamma^{\mu\nu} = h^{\mu\nu} - c^{-1}u^\mu u^\nu$, $h^{\mu\nu}$ the Galilean metric and u^μ a constant vector field s.th. $u^\mu\tau_\mu = 1$, one has the Maxwell equations in a "Galilean" (not Galilean-invariant) version with u^μ describing the rest system of the aether (see [43]). One then looks at hypersurfaces along which singularities can propagate. The result is that the conormal n_μ of such surfaces has to satisfy $P(n) = 0$. Put differently, one can look at the 8 x 6 – principal symbol of the Maxwell equations: then $P(k) = 0$ is exactly the condition for this principal symbol to have nontrivial kernel. If one considered an appropriately chosen subset amongst (3), the evolution equations, one would obtain an equation of the form (1), whose characteristic polynomial contains $P(k)$ as a factor. We will treat the vacuum case of this later.

Our last and most complicated example comes from elasticity [6]:

Example 5. Take $n = 4, l = 2, m = 3$ in (1) with

$$M^{\mu\nu}_{AB} = -G_{AB}u^\mu u^\nu + C^{\mu\nu}_{AB}\ , \tag{4}$$

where $G_{AB} = G_{(AB)}$ and $C^{\mu\nu}_{AB} = C^{\nu\mu}_{BA}$ and $C^{\mu\nu}_{AB}\tau_\nu = 0$ for some covector τ satisfying $(u, \tau) = u^\mu\tau_\mu = 1$. The theory is intrinsically quasilinear: all quantities entering (4) are functions of f and ∂f and in general also of x. For example f^A is required to have maximal rank, and u^μ satisfies $u^\mu(\partial_\mu f^A) = 0$. Furthermore $C^{\mu\nu}_{AB} = C_{ADBE}(\partial_\rho f^D)(\partial_\sigma f^E)h^{\rho\mu}h^{\sigma\nu}$, with $h^{\mu\nu}, \tau_\mu$ being, in the nonrelativistic case, the absolute Galilean objects, or, in the relativistic case, $h^{\mu\nu} = g^{\mu\nu} + u^\mu u^\nu$ and $\tau_\mu = -g_{\mu\nu}u^\nu$.

There are the following basic constitutive assumptions.

$$G_{AB} \text{ is positive definite}, C^{\mu\nu}_{AB}m^Am^B\eta_\mu\eta_\nu > 0 \text{ for } m \neq 0,\ \eta \wedge \tau \neq 0 \tag{5}$$

Defining the linear map $(\mathbf{M})^A{}_B$ by $(\mathbf{M})^A{}_B(k) = -(u,k)^2\delta^A{}_B + G^{AD}C^{\mu\nu}_{DB}k_\mu k_\nu$, the polynomial $P(k)$ can, by general linear algebra, be written as

$$6\ P(k) = (tr\mathbf{M})^3 - 3\ (tr\mathbf{M^2})(tr\mathbf{M}) + 2\ tr\mathbf{M^3}\ . \tag{6}$$

[3]These equations play a certain role in current searches for violations of Lorentz invariance in electrodynamics [30]

It will follow from a more general result, to be shown below, that the 6th order polynomial $P(k)$ is hyperbolic with respect to ξ_μ in some neighbourhood of τ_μ. In the special case of an isotropic solid, the "elasticity tensor" C_{ABDE} has to be of the form

$$C_{ABDE} = l\ G_{AB}G_{DE} + 2m\ G_{D(A}G_{B)E}\ , \tag{7}$$

and the second of (5) is satisfied iff $c_2^2 = m > 0$, $c_1^2 = l + 2m > 0$. The polynomial $P(k)$ turns out to reduce to the form

$$P(k) \sim (\gamma_1^{\mu\nu}k_\mu k_\nu)(\gamma_2^{\rho\sigma}k_\rho k_\sigma)^2 \tag{8}$$

with $\gamma_{1,2}^{\mu\nu} = h^{\mu\nu} - c_{1,2}^{-2}\ u^\mu u^\nu$. The quantities c_1 and c_2 are the phase velocities of pressure and shear waves respectively. If the medium is elastically anisotropic, such as a crystal, one can start by classifying possible fourth-rank tensors C_{ABDE} according to the symmetry group of the crystal lattice, allow for dislocations, etc. The richness of possible structure of $\mathcal{C}^*$ and the corresponding range of captured physical phenomena – studied by theoreticians and experimentalists – is enormous.

This ends our list of examples. We now turn to some general properties of hyperbolic polynomials and their physical interpretation. It is clear from the definition that $\mathcal{C}^*$ has codimension 1: since $P(\eta+\lambda\xi)$ has to have at least one complex root for each η, and the roots are all real, there is at least one real root. And since there are no more than $p \cdot l$ different roots, $\mathcal{C}^*$ can not have larger codimension. It is then known from real algebraic geometry that $\mathcal{C}^*$ consists of smooth hypersurfaces outside a set of at least codimension 2 (see [9]). The roots $\lambda_i(\xi,\eta)$ can for fixed ξ be assumed to be ordered according to $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_p$ for all η. The set of points $k = \eta + \lambda_i(\xi,\eta)\xi$ is called the i'th sheet of $\mathcal{C}^*$. The hypersurface $\mathcal{C}^*$ has to be smooth at all points k lying on a line intersecting p different sheets[4]. In particular all sheets are everywhere smooth when P is strictly hyperbolic.

Next recall that the defining property of a hyperbolic polynomial refers to a particular covector ξ. That covector however is not unique. It is contained in a unique connected, open, convex, positive cone $\Gamma^*(\xi)$ of covectors ξ' sharing with ξ the property that $P(\xi') \neq 0$ and $P(\eta+\lambda\xi')$ has only real zeros $\lambda_i(\xi',\eta)$ [19]. Note that $\Gamma^*(\xi) = -\Gamma^*(-\xi)$. Furthermore $\partial\Gamma^*(\xi) \subset \mathcal{C}^*$, and $\Gamma^*(\xi)$ is that connected component of the complement of $\mathcal{C}^*$ containing ξ. Not all points of $\partial\Gamma^*(\xi)$ have to be smooth points of $\mathcal{C}^*$.

The roots $\lambda_i(\xi,\eta), i = 1,\ldots,p$, due to the homogeneity of P, are homogenous in ξ of order -1 and positively homogenous of order 1 as a function of η. They also satisfy $\lambda_i(\xi,\eta) = -\lambda_{p+1-i}(\xi,\eta)$. At regular points of $\mathcal{C}^*$, i.e. when the gradient of P at $\eta + \lambda_i\xi$ is non-zero, $\lambda_i(\xi,\eta)$ is a smooth function

[4]The reason is that a polynomial of order p in one real variable, if it has p different zeros, has non-vanishing derivative at each zero, so k is a non-critical point of P.

of its arguments due to the implicit function theorem. Next choose a vector $X \in \mathbb{R}^n$ so that $(X, \xi') > 0$ for all $\xi' \in \Gamma^*(\xi)$. We call such a vector "causal". We now look at the intersection S of the hyperplane $(X, \xi') = 1$ with $\mathcal{C}^*$. Note $(X, \xi') = 1$ is transversal to $\mathcal{C}^*$ at smooth points of $\mathcal{C}^*$, so S is smooth there also. Note also that S may be empty, as in Example 1. For reasons explained below, S is often called "slowness surface". To describe S more concretely, we pick some $\tau \in \Gamma^*(\xi)$ with $(X, \tau) = 1$. The pair X, τ constitutes a "rest frame". Using it, we can decompose every covector k as $k_\mu = \tau_\mu + k^\perp_\mu$ where $k^\perp$ is tangential to the hyperplane, i.e. $(X, k^\perp) = 0$. This k lies on $\mathcal{C}^*$ iff $\lambda_i(\tau, k^\perp) = 1$ for some i. Thus S consists of the sheets $\lambda_J(\tau, k^\perp) = 1$, viewed as $(n-2)$-surfaces in $k^\perp \in \mathbb{R}^{n-1}$. Here J runs through some subset of the i's parametrizing λ_i from above. Clearly, as J increases, these sheets form a nested family of not necessarily compact surfaces[5]. The innermost of these surfaces is nothing but the intersection of $\partial\Gamma^*(\xi)$ with the hyperplane $(X, \xi') = 1$ and is hence convex. We call $\mathcal{C}^*(\xi)$ those components of $\mathcal{C}^*$ which consist of half-rays connecting the origin with the points of S. In the examples 2, 4, 5 the set S and $\mathcal{C}^*(\xi)$ consist of at most 1, respectively 2 and 3 sheets. For the last-mentioned case, see [15]. Not all the occurring sheets are compact. It is possible for example for $P(k)$ to be an irreducible hyperbolic polynomial with some sheets of S compact and others non-compact: this is the case e.g. for the acoustic modes in magnetohydrodynamics[13].

We now explain the name "slowness surface". Consider the hyperplane in $\mathbb{R}^n$ given by $(x, k) = 0$ for fixed $k \in \mathcal{C}^*$, i.e. the wave front of the plane wave associated with k. To measure the "speed" at which this wave front moves, decompose observers with tangent V which "move with this wave front", i.e. such that $(V, k) = 0$, according to $V = X + v^\perp$. It follows that $(v^\perp, k^\perp) = -1$. Thus, if there is a natural "spatial" metric h mapping elements $l^\perp$ into elements $w^\perp = h \circ l^\perp$ orthogonal to τ, one can define the "phase velocity" $v^\perp_{\mathrm{ph}} = -\|k^\perp\|^{-2}\, h \circ k^\perp$. Thus, the smaller $k^\perp$ is, the larger the phase velocity. Of course the equation $(v^\perp, k^\perp) = -1$ does not define $v^\perp$ uniquely. But there is a "correct" choice for V tangential to the wave front, called "ray or group velocity", which is independent of any spatial metric, and which is defined at least when k is a smooth point of $\mathcal{C}^*$: this V is given by the conormal to $\mathcal{C}^* \subset (\mathbb{R}^n)^*$ at k, which by duality is a vector $\in \mathbb{R}^n$. If k is in addition a non-critical point, this ray velocity V^μ is $\sim \partial/\partial k_\mu P(k)$, which satisfies $k_\mu V^\mu = 0$ by the positive homogeneity of $P(k)$. The spatial group velocity in the frame X, τ can then be written as which is also the textbook expression.

$$(v^\perp_{\mathrm{gr}})^\mu(k^\perp) = \left(\tau_\lambda \frac{\partial P}{\partial k_\lambda}\right)^{-1} \left(\delta^\mu{}_\nu - X^\mu \tau_\nu\right) \left.\frac{\partial P}{\partial k_\nu}\right|_{k=\tau+k^\perp}, \tag{9}$$

We should add a cautionary remark here. Although the differential topology of the slowness surface is independent of the choice of X satisfying

[5] In particular, when sheets seem to pass through each other, the two sides are counted as belonging to different sheets.

$(X, \xi') > 0$ for all $\xi' \in \Gamma^*(\xi)$, its detailed appearance, and physical quantities such as phase velocity, group velocity or angle between two rays do of course depend on the choice of rest system X, τ and a notion of spatial metric with respect to that observer. Of course there will be, for any particular physical theory, a singled-out class of rest systems, e.g. τ_μ can be the absolute object in a Galilean spacetime or be of the form $\tau_\mu = -g_{\mu\nu}X^\nu$ in a relativistic theory. Or the slowness surface can have more symmetry (say symmetry with respect to reflection at the origin) in some rest system than in others, as is the case with crystal optics or elasticity. For a careful discussion of these issues, in the more specialized context of "ray-optical structures" on a Lorentzian spacetime, consult [34].

We now come back to the "ray" concept. If k is a smooth critical point of $\mathcal{C}^*$, finding the map $k \mapsto V(k)$ is already a nontrivial problem in algebraic geometry [35]. If k is not a smooth point of $\mathcal{C}^*$, there is no unique assignment of a group velocity to k. Still well-defined is the set $\mathcal{C}$ of all $V \neq 0$ satisfying

$$(V, k) = 0 \qquad \text{where} \quad P(k) = 0 \,, \tag{10}$$

called the dual or ray cone. Loosely speaking, each sheet of the ray cone corresponds to a spherical wave front tangent to (or "supported by") the planar wave fronts defined by the different points k in some corresponding sheet of $\mathcal{C}^*$ [13]. There holds $(\mathcal{C}^*)^* = \mathcal{C}$. The dual cone is again an algebraic cone, which, except in degenerate cases, is again the zero-set of a single homogenous polynomial. The structure of this dual cone, in particular its singularity structure which can be very complicated, is another difficult matter of real algebraic geometry. For example the degree of its defining polynomial is in general much higher than that of $\mathcal{C}^*$ (see [37],[21]). This "dual" polynomial need not be hyperbolic: in order to be hyperbolic it would have to have a central sheet which is convex, which is not the case for some of the examples one finds in the literature. In our examples from above the situation is as follows: In our Example 1 the dual cone $\mathcal{C}^*$ consist of the two halflines $\{\alpha a^\mu | \alpha > 0\}$ and $\{\alpha a^\mu | \alpha < 0\}$. The cone dual to the quadratic cone $g^{\mu\nu}k_\mu k_\nu = 0$ in Example 2 is given by $g_{\mu\nu}V^\mu V^\nu = 0$ with $g_{\mu\nu}g^{\nu\lambda} = \delta_\mu{}^\lambda$. For a nonrelativistic acoustic cone $\gamma^{\mu\nu} = h^{\mu\nu} - (1/c^2)u^\mu u^\nu$ we obtain for the ray cone $\gamma_{\mu\nu} = h_{\mu\nu} - c^2\tau_\mu\tau_\nu$, where $h_{\mu\nu}$ is the unique tensor defined by $h_{\mu\nu}u^\nu = 0$ and $h_{\mu\nu}h^{\nu\lambda} = \delta_\mu{}^\lambda - \tau_\mu u^\lambda$. If one has two sound cones, as in isotropic elasticity, it is the faster ray cone which lies outside. In Example 3 $\mathcal{C}$ is given as a subset of vectors X^μ in a linear space T, which is the annihilator of the null space of $s^{\mu\nu}$, namely where this subset is given by $s_{\mu\nu}X^\mu X^\nu = 0$, where $s_{\mu\nu}$ is the inverse of $s^{\mu\nu}$ on T. In the magnetohydrodynamic example the preceding statement corresponds to the fact that Alfvén waves "travel along the direction of the magnetic field". For Example 4 the ray cone $\mathcal{C}$ is a 4th order cone of the same type as $\mathcal{C}^*$, a fact already known by Ampère in the case of crystal optics and shown generally in [36]. For anisotropic elasticity the structure of the ray cone does not seem to be fully known, except for a general upper

bound on its degree, namely 150 on grounds of general algebraic geometry (see [15],[37],[21]) and detailed studies for certain specific crystal symmetries – which give rise to a beautiful variety of acoustic phenomena [44][6].

3 Initial Value Problem

We now come to the issue of posing an initial value problem for hyperbolic equations of the form of (1). This requires two things: firstly a notion of "spacelike" initial value surface, secondly a notion of domain of dependance. Not surprisingly these notions can be formulated purely in terms of the characteristic polynomial. A hypersurface Σ in $\mathbb{R}^n$ will be called spacelike, if it has a conormal n_μ lying everywhere in $\Gamma^*(\xi)$ for some ξ. If the equation (1) is nonlinear, every property concerning the characteristic polynomial has to refer to the data of some reference field f_0, i.e. the value of f_0 on Σ and those of its derivatives up to order $l-1$. It is then the case that Σ is spacelike also for any sufficiently near-by data. The reason is that $\xi' \in \Gamma^*(\xi)$ can be characterized by $\lambda_1(\xi, \xi') > 0$, and the eigenvalues λ_i, being zeros of a polynomial having real roots only, depend continuously on the coefficients of this polynomial [1]. A point x in $\mathbb{R}^n$ is said to lie in the domain of dependence of Σ if each causal curve (i.e. each curve whose tangent vector X satisfies $(X, \xi') \neq 0$ for all $\xi' \in \Gamma^*(\xi)$) through x which is inextendible intersects Σ exactly once. The Cauchy problem for (1) is said to be well-posed if, for the above data, there is a unique solution in some domain of dependence of Σ and, secondly, if this solution depends in some appropriate sense continuously on the data. The question then is whether well-posedness holds under the above conditions. The answer is affirmative when (1) is linear with constant coefficients and the lower-order terms are absent. Then the initial value problem can be solved "explicitly" by using a fundamental solution ("Green function" in the physics literature) – which in turn can be obtained e.g. by the Fourier transform. By a refined version of a well-known argument in physics texts concerning the wave equation in Minkowski space (see e.g. [3]), one can show that the fundamental solution is supported in $\Gamma(\xi)$, which is the closure of the set of causal vectors just described. The set $\Gamma(\xi)$ is a closed, convex cone, dual to $\Gamma^*(\xi)$. If the outermost component of the cone $\mathcal{C}(\xi)$ dual to $\mathcal{C}^*(\xi)$ is convex, its closure is the same as $\Gamma(\xi)$, otherwise its convex closure is the same as $\Gamma(\xi)$. If one is interested in finer details than just wellposedness, even the linear, constant-coefficient case becomes very nontrivial. An example is the question of the existence of "lacunas", i.e. regions in $\Gamma(\xi)$ where the fundamental solution vanishes. For isotropic elasticity mentioned in Example 4, when $c_2 < c_1$ (which is the experimentally relevant case), the fundamental solution vanishes inside the inner shear cone determined by c_2.

[6] There are computer codes designed for algebraic elimination, which might be worth applying to this problem [23].

(Note that "inner" and "outer" are interchanged under transition between normal and ray cone.) For anisotropic elasticity this issue, or the somewhat related question of the detailed time decay, already presents great difficulties (see [15, 39]). The existence of lacunas for general, linear hyperbolic systems with constant coefficients was studied in [3].

The problem now is that many field equations in physics give rise to variable coefficients, to various forms of lower-order terms and-or nonlinearities. But if one has a system of PDE's with hyperbolic characteristic polynomial (such systems are often called "weakly hyperbolic") , which in addition has a well posed initial value problem, a perturbation of the coefficients will in general destroy the latter property (see e.g. [31]). It is thus hard to get any further without additional assumptions. One such assumption is that of having a symmetric hyperbolic system. This is given by a system of the form of (1) with $l = 1$. It is furthermore assumed that

$$M_{AB}^{\ \ \mu} = M_{(AB)}^{\ \ \ \mu} \tag{11}$$

and that there exists ξ_μ so that

$$M_{AB}(\xi) = M_{AB}^{\ \ \mu}\ \xi_\mu \quad \text{is positive definite .} \tag{12}$$

The symmetric hyperbolic system has a characteristic polynomial which is hyperbolic with respect to ξ. To see this one simply observes that the equation

$$det(M_{AB}^{\ \ \mu}(\eta_\mu + \lambda\xi_\mu)) = 0 \tag{13}$$

characterizes eigenvalues of the quadratic form $M_{AB}(\eta)$ relative to the metric $M_{AB}(\xi)$ – and these eigenvalues have to be real. There is then, for quasilinear symmetric hyperbolic systems, a rigorous existence statement along the lines informally outlined at the beginning of this section [29]. The uniqueness part uses the concept of "lens-shaped domains" (see e.g. [18]) which is essentially equivalent to that of domain of dependence above.

Several field theories of physical importance naturally give rise to a symmetric hyperbolic system. An example is afforded by the hydrodynamics of a perfect fluid both nonrelativistically and relativistically[7]. The most prominent examples are perhaps the Maxwell equations in vacuo and the vacuum Bianchi identities in the Einstein theory. For the latter this was first observed in [17]. For completeness we outline a proof for the well-known Maxwell case following [45]. We have that

$$\nabla^\nu F_{\mu\nu} = 0, \qquad \nabla_{[\mu} F_{\nu\lambda]} = 0 \tag{14}$$

with ∇_μ being the covariant derivative with respect to $g_{\mu\nu}$, a Lorentz metric on $\mathbb{R}^4$. These are 8 equations for the 6 unknowns $F_{\mu\nu}$. Next pick a timelike vector field u^μ with $u^2 = -1$ and define electric and magnetic fields by

[7] For an elegant treatment of the latter, see [16]

$$E_\mu = F_{\mu\nu}u^\nu, \qquad B_{\mu\nu\lambda} = 3F_{[\mu\nu}u_{\lambda]} \,, \tag{15}$$

so that

$$F_{\mu\nu} = -2E_{[\mu}u_{\nu]} - B_{\mu\nu\lambda}u^\lambda \,. \tag{16}$$

We assume for simplicity that u^μ is covariant constant, otherwise the ensuing equations contain zero'th order terms which are of no concern to us. The operator

$$\nabla_{\mu\nu} = 2u_{[\mu}\nabla_{\nu]} \tag{17}$$

contains derivatives only in directions orthogonal to u^μ. Using Eq.'s (14) we find the evolution equations

$$3\nabla_{[\mu\nu}E_{\lambda]} = -u^\rho\nabla_\rho B_{\mu\nu\lambda}, \qquad \nabla^{\lambda\rho}B_{\nu\lambda\rho} = 2u^\rho\nabla_\rho E_\nu \,. \tag{18}$$

Taking now $u^\lambda\nabla_\rho$ of (15), we rewrite the evolution equations in the form

$$W_{\mu\nu}^{\mu'\nu'\lambda}\nabla_\lambda F_{\mu'\nu'} = 0 \,. \tag{19}$$

Now take the positive definite metric

$$w^{\mu\nu} = 2u^\mu u^\nu + g^{\mu\nu} \,. \tag{20}$$

Consider now the positive definite metric $a^{\mu\nu\lambda\rho} = 2w^{\rho[\mu}w^{\nu]\lambda}$ on the space of 2-forms and use it to raise indices in $W_{\mu\nu}^{\mu'\nu'\lambda}$: One obtains quantities $W^{\mu\nu\mu'\nu'\lambda}$ satisfying

$$W^{\mu\nu\mu'\nu'\lambda} = W^{\mu'\nu'\mu\nu\lambda}, \qquad W^{\mu\nu\mu'\nu'\lambda}u_\lambda \sim a^{\mu\nu\mu'\nu'} \,. \tag{21}$$

Thus the (18) are symmetric hyperbolic with respect to u_μ. For the characteristic polynomial one finds

$$P(k) \sim (u^\mu k_\mu)^2(g^{\rho\nu}k_\rho k_\nu)^2 \,. \tag{22}$$

We now turn to 2nd order equations. Let us assume that the quantities $M_{AB}^{\mu\nu}$ of (1) satisfy

$$M_{AB}^{\mu\nu} = M_{BA}^{\nu\mu} \,. \tag{23}$$

This is necessarily the case when (1) comes from a variational principle, because then

$$M_{AB}^{\mu\nu} \sim \frac{\partial^2\mathcal{L}}{(\partial\partial_\mu f^A)(\partial\partial_\nu f^B)} \,, \tag{24}$$

where $\mathcal{L} = \mathcal{L}(x, f, \partial f)$. Of course, since only the quantities $M_{AB}^{(\mu\nu)}$ enter the differential equation (1), one might as well have assumed the stronger conditions $M_{AB}^{\mu\nu} = M_{AB}^{\nu\mu} = M_{BA}^{\mu\nu}$. This is not usually done in continuum mechanics. The reason is that, while the PDE (and hence the characteristic polynomial) is unaffected by the above symmetrization, other physical quantities in the

theory, like the stress, depend on the unsymmetrized object – and such objects typically enter, if not the equation, then the natural boundary conditions for the equation on the surface say of an elastic body (see [8]). A related reason is that the symmetrized object would hide other symmetries – present in some situations – which are more fundamental such as invariances under isometries. For example the object $C_{A(B|D|E)}$, with C_{ABDE} the elasticity tensor of (7), is not symmetric in (AB) and (DE). But the latter symmetry is important for understanding the solutions of the linearized equations of motion when the spacetime has Killing vectors. The work [10] also uses the unsymmetrized form of $M^{\mu\nu}_{AB}$, the reason being that in this approach one is only interested in properties of $M^{\mu\nu}_{AB}$ which do not change when a total divergence is added to the Lagrangian, and the stronger symmetry, if present, would in general be destroyed by such an addition. Next it is assumed that there exists a pair X^μ, ξ_ν satisfying

$$M^{\mu\nu}_{AB}\xi_\mu\xi_\nu \text{ is negative definite} \tag{25}$$

and

$$M^{\mu\nu}_{AB}(m^A\eta_\mu)(m^B\eta_\nu) > 0 \text{ for all } m^A\eta_\mu \neq 0 \text{ with } (X,\eta) = 0\,. \tag{26}$$

The conditions (25,26) essentially state that the PDE is the sum of a "timelike part" and an "elliptic part", the latter obeying the Legendre-Hadamard condition of the calculus of variations [22]. If the equation (1) has $l = 2$ and satisfies (23,25,26), the system is called regular hyperbolic with respect to ξ. We now check that every regular hyperbolic system with respect to ξ is weakly hyperbolic with respect to ξ. The characteristic condition reads

$$det(M^{\mu\nu}_{AB}(\eta_\mu + \lambda\xi_\mu)(\eta_\nu + \lambda\xi_\nu)) = 0 \tag{27}$$

The covector η in (27) can be decomposed as $\eta = \frac{(X,\eta)}{(X,\xi)}\xi + l$ where l satisfies $(X, l) = 0$. Thus we can after redefining λ assume that η in (27) has $(X,\eta) = 0$. Defining $G_{AB} = -M_{AB}(\xi) = -M^{\mu\nu}_{AB}\xi_\mu\xi_\nu$, $V_{AB} = M_{AB}(\eta)$, $Q_{AB} = M^{\mu\nu}_{(AB)}\xi_\mu\xi_\nu$, consider the eigenvalue problem

$$\mathcal{D}\hat{f} = \lambda\mathcal{E}\hat{f}\,, \tag{28}$$

in

$$\hat{f} = \begin{pmatrix} u^A \\ v^B \end{pmatrix}$$

where the quadratic forms $\mathcal{D}$, $\mathcal{E}$ are given by

$$\mathcal{D} = \begin{pmatrix} 0 & V_{AB} \\ V_{AB} & 2Q_{AB} \end{pmatrix}$$

and

$$\mathcal{E} = \begin{pmatrix} V_{AB} & 0 \\ 0 & G_{AB} \end{pmatrix}.$$

Since $\mathcal{E}$ is positive definite, all eigenvalues λ are real. But (28) for $\hat{f} \neq 0$ is equivalent to

$$(-G_{AB}\lambda^2 + 2Q_{AB}\lambda + V_{AB})v^B = 0\,, \tag{29}$$

for $v^A \neq 0$ which in turn is equivalent to (27). This proves our assertion that regular hyperbolic systems are weakly hyperbolic. (Note that every "timelike vector" X in the sense of (26) is causal, i.e. $(X, \xi) \neq 0$ for all $\xi \in \Gamma^*(\xi)$, but not conversely.) We can now come back to Example 5. The leading-order coefficients $M^{\mu\nu}_{AB}$ in (4) clearly belong to a regular hyperbolic system, when we choose the vector $X^\mu \sim u^\mu$. It then follows from the preceding result that the polynomial in (6) is indeed a hyperbolic polynomial.

As with symmetric hyperbolic systems, it turns out that there is, for regular hyperbolic systems, a local existence theorem [27] along the lines sketched at the beginning of this section. The appropriate domain of dependence theorem is proved in [10].

One may ask the question if it is possible to convert a regular hyperbolic system into an equivalent symmetric hyperbolic one by introducing first derivatives as additional dependent variables (at the price of course of having to solve constraints for the initial data). (This was the approach we originally followed for elasticity in [6], since we were unaware that there was already an existence theorem which applied, namely [27]). If the condition (12) is provisionally ignored, it turns out this is possible provided that $M^{\mu\nu}_{AB}$ is of the form of (4) for some pair u^μ, τ_ν, i.e. certain cross-terms vanish[8]. But the positivity condition (12) will not always be satisfied. (Essentially this requires the "rank-one positivity" condition (26) to be replaced by the stronger rank-two positivity: $M^{\mu\nu}_{AB} m^A{}_\mu m^B{}_\nu > 0$ for all $m^A{}_\mu \neq 0$ with $X^\mu m^A{}_\mu = 0$.) In the case of isotropic elasticity it was shown in [6] that one can add to $M^{\mu\nu}_{AB}$ a term of the form $\Lambda^{\mu\nu}_{AB}$, which has the symmetries $\Lambda^{\mu\nu}_{AB} = \Lambda^{[\mu\nu]}_{[AB]}$, so that both the field equations and the requirement (23) remains unchanged, but at the same time condition (12) is valid. However it is an algebraic fact that such a trick does not always work (see [42, 38]).

Finally let us mention the notion of strong hyperbolicity, which is intermediate between weak hyperbolicity and symmetric or regular hyperbolicity in the first or second order case respectively. This notion, which involves the tool of pseudodifferential reduction [40, 41], also gives wellposedness but has greater flexibility, see [33] for applications to the Einstein equations. It would be interesting to see if the chain "weakly hyperbolic – strongly hyperbolic – symmetric or regular hyperbolic" has an analogue for PDE's of order greater than 2.

[8] In [7] I claimed this to be possible even without these cross-terms vanishing. I now see I have no proof of this assertion.

Acknowledgements

I am indebted to Domenico Giulini for carefully reading the manuscript and suggesting several corrections and improvements. I also would like to thank Helmuth Urbantke for several very helpful conversations and Piotr T. Chruściel and Helmut Rumpf for useful comments. This work was supported by Fonds zur Förderung der Wissenschaftlichen Forschung (Projekt P16745-N02)

References

1. D. Alekseevsky, A. Kriegl, M. Losik, P.W. Michor: Choosing roots of polynomials smoothly. Israel J. Math. **105**, 203–233 (1998) 109
2. A.M. Anile: *Relativistic Fluids and Magneto-Fluids* (Cambride University Press, Cambridge 1989) 104
3. M.F. Atiyah , R. Bott, L. Gårding: Lacunas for hyperbolic differential operators with constant coefficients I; II. Acta Math. **124**, 109–189 (1970); **131**, 145–206 (1973) 109, 110
4. S. Barcelo, S. Liberati, S. Sonego, M. Visser: Causal structure of acoustic spacetimes (2004) `gr-qc/04080221` 104
5. H.H. Bauschke, O. Güler, A.S. Lewis, H.S. Sendov: Hyperbolic polynomials and convex analysis. Can. J. Math. **53**, 470–488 (2001) 104
6. R. Beig, B.G. Schmidt: Relativistic elasticity. Class. Quantum Grav. **20**, 889–904 (2003) 105, 113
7. R. Beig: Nonrelativistic and relativistic continuum mechanics. To appear in the Proceedings of the Seventh Hungarian Relativity Workshop, Sarospatak, Hungary. `gr-qc/0403073` 113
8. R. Beig, M. Wernig-Pichler: The Cauchy problem for relativistic elastic bodies with natural boundary conditions. In preparation (2004) 112
9. R. Benedetti, J.-J. Risler: *Real Algebraic and Semialgebraic Sets* (Hermann, Paris 1990) 106
10. D. Christodoulou: *The Action Principle and Partial Differential Equations* (Princeton University Press, Princeton 2000) 102, 112, 113
11. D. Christodoulou: On hyperbolicity. Contemporary Mathematics **283**, 17–28 (2000) 102
12. P.T. Chruściel: Black holes. In *Conformal Structure of Spacetime*, ed by J. Frauendiener and H. Friedrich, Lecture Notes in Physics **604** (Springer, Heidelberg 2002) 104
13. R. Courant, D. Hilbert: *Methods of Mathematical Physics, Volume II* (Interscience Publishers, New York 1962) 102, 107, 108
14. D.M. DeTurck, D. Yang: Existence of elastic deformations with prescribed principal strains and triply orthogonal systems. Duke Math. Journal **51**, 243–260 (1984) 103
15. G.F.D. Duff: The Cauchy problem for elastic waves in an anisotropic medium. Transactions Roy. Soc. A **252**, 249–273 (1960) 107, 109, 110
16. J. Frauendiener: A note on the relativistic Euler equations. Class. Quantum Grav. **20**, L193–L196 (2003) 110

17. H. Friedrich: On the regular and the asymptotic characteristic initial value problem for Einstein's vacuum field equations. Proc. Roy. Soc. A **375**, 169–184 (1985) 110
18. H. Friedrich, A. Rendall: The Cauchy problem for the Einstein equations. In: *Einstein's Field Equations and their Physical Interpretation*, ed by B. G. Schmidt, Lecture Notes in Physics **540** (Springer, Heidelberg 2000) 101, 110
19. L. Gårding: An inequality for hyperbolic polynomials. J. of Mathematics and Mechanics **8**, 957–965 (1959) 104, 106
20. L. Gårding: History of the mathematics of double refraction. Arch. Hist. Ex. Sci. **40**, 355–385 (1989) 101
21. I.M. Gelfand, M.M. Kapranov, A.V. Zelevinsky: *Discriminants, Resultants and Multidemensional Determinants* (Birkhäuser, Boston 1994) 108, 109
22. M. Giaquinta, S. Hildebrandt: *Calculus of Variations II* (Springer, Berlin 1996) 112
23. G.-M. Greuel, G. Pfister, H. Schönemann: SINGULAR 2.0. A computer algebra system for polynomial computations. Centre for Computer Algebra, University of Kaiserslautern (2001) `http://www.singular.uni-kl.de` 109
24. F.W. Hehl, Y.N. Obukhov: *Foundations of Classical Electrodynamics – Charge, Flux, and Metric* (Birkhäuser, Basel 2003) 105
25. G. Herglotz: *Vorlesungen über die Mechanik der Kontinua*. Göttingen lectures of 1926 and 1931, elaborated by R. B. Guenther and H. Schwerdtfeger (Teubner, Leipzig 1988) 104
26. L. Hörmander: Hyperbolic systems with double characteristics. Comm. Pure Appl. Math. **46**, 89–106 (1993) 103
27. T.J.R. Hughes, T. Kato, J.E. Marsden: Well-posed quasilinear second-order hyperbolic systems with applications to Nonlinear Elastodynamics and General Relativity. Arch. Rat. Mech. Anal. **63**, 273–294 (1977) 113
28. F. John: Algebraic conditions for hyperbolicity of systems of partial differential equations. Comm.Pure Appl.Math. **31**, 89–106 (1978) 103
29. T. Kato: The Cauchy problem for quasi-linear symmetric hyperbolic systems. Arch. Rat. Mech. Anal. **58**, 181–205 (1975) 110
30. A. Kostelecký, M. Mewes: Signals for Lorentz violation in electrodynamics. Phys. Rev. D **66**, 056005–056028 (2002) 105
31. H.-O. Kreiss, O. Ortiz: Some mathematical and numerical questions connected with first and second order time dependent systems of partial differential equations. In: *The Conformal Structure of Space-Time*, ed by J. Frauendiener and H. Friedrich), Lecture Notes in Physics **604** (2002) 110
32. C. Lämmerzahl, F. W. Hehl: Riemannian cone from vanishing birefringence in premetric electrodynamics. (2004) `gr-qc/0409072` 105
33. G. Nagy, O.E. Ortiz, O.A. Reula: Strongly hyperbolic second order Einstein's evolution equations. Phys. Rev. D **70**, 044012(15) (2004) 113
34. V. Perlick: *Ray Optics, Fermat's Principle, and Applications to General Relativity*, Lecture Notes in Physics **m61** (Springer, Heidelberg 2000) 108
35. J. Rauch: Group velocity at smooth points of hyperbolic characteristic varieties. Astérisque **284**, 265–269 (2003) 108
36. G.F. Rubilar: Linear pre-metric electrodynamics and deduction of the light-cone. Ann. Phys. (Leipzig) **11**, 717–782 (2002) 108
37. G. Salmon: *Lectures Introductory to the Modern Higher Algebra* (Dublin 1885; reprinted by Chelsea Publ., New York 1969) 108, 109
38. D. Serre: Formes quadratiques et calcul des variations. J. Math. Pures et Appl. **62**, 177–196 (1983) 113

39. M. Stoth: Decay estimates for solutions of linear elasticity for anisotropic media. Math. Meth. Appl. Sci. **19**, 15–31 (1996) 110
40. M.E. Taylor: *Pseudodifferential Operators and Nonlinear PDE* (Birkhäuser, Boston 1991) 113
41. M.E. Taylor: *Pseudodifferential Operators* (Princeton University Press, Princeton 1981) 113
42. F.J. Terpstra: Die Darstellung biquadratischer Formen als Summen von Quadraten mit Anwendung auf die Variationsrechnung. Math. Ann. **116**, 166–180 (1938) 113
43. A. Trautman: Comparison of Newtonian and relativistic theories of space-time. In: *Perspectives in Geometry and Relativity*, ed by B. Hoffmann, (Indiana University Press 1966) 105
44. J.P. Wolfe: *Imaging Phonons* (Cambridge University Press, Cambridge 1998) 109
45. A.Ç. Zenginoglŭ: Ideal magnetohydrodynamics in curved spacetime. Masters Thesis, University of Vienna (2003) 104, 110

Elliptic Systems

Sergio Dain

Max-Planck-Institut für Gravitationsphysik, Am Mühlenberg 1, 14476 Golm, Germany
dain@aei.mpg.de

Abstract. In this contribution I will review some basic results on elliptic boundary value problems with applications to General Relativity.

1 Introduction

Elliptic problems appear naturally in physics mainly in two situations: as equations which describe equilibrium (for example, stationary solutions in General Relativity) and as constraints for the evolution equations (for example, constraint equations in Electromagnetism and General Relativity). In addition, in General Relativity they appear often as gauge conditions for the evolution equations.

The model for all elliptic equations is the Laplace equation. Let us consider the Dirichlet boundary-value problem for this equation

$$\Delta u = f \text{ on } \Omega, \quad u = g \text{ on } \partial\Omega \,, \tag{1}$$

where Ω is a bounded, smooth, domain in $\mathbb{R}^n$ with boundary $\partial\Omega$; f, g are smooth functions and Δ is the Laplace operator in $\mathbb{R}^n$.

It is a well known result that for every source f and every boundary value g there exist a unique, smooth, solution u of (1). We would like to generalize equations (1) for more general operators and more general boundary conditions.

The first step in this generalization is given by the Neumann problem

$$\Delta u = f \text{ on } \Omega, \quad n^i \partial_i u = 0 \text{ on } \partial\Omega \,, \tag{2}$$

where n^i is the outward unit normal to $\partial\Omega$, the index i takes values $i = 1, \ldots, n$ and ∂_i denotes partial derivative with respect to the $\mathbb{R}^n$ coordinate x_i.

There exist two main differences between the Neumann and the Dirichlet problem: (i) The solution to the Neumann problem is not unique, for a given solution we can add a constant and obtain a new solution. Moreover, the constants are the only solutions of the homogeneous problem

$$\Delta u = 0 \text{ on } \Omega, \quad n^i \partial_i u = 0 \text{ on } \partial\Omega \,. \tag{3}$$

S. Dain: *Elliptic Systems*, Lect. Notes Phys. **692**, 117–139 (2006)
www.springerlink.com

To see this, we multiply (3) by u and use the divergence theorem

$$0 = \int_\Omega u\Delta u = \int_\Omega \partial^i(u\partial_i u) - \partial^i u\partial_i u \tag{4}$$

$$= \oint_{\partial\Omega} un^i\partial_i u - \int_\Omega \partial^i u\partial_i u \tag{5}$$

$$= -\int_\Omega \partial^i u\partial_i u\,. \tag{6}$$

(ii) The source f cannot be arbitrary. We integrate in Ω equation (2) to obtain a necessary condition for f

$$0 = \oint_{\partial\Omega} n^i\partial_i u = \int_\Omega \Delta u = \int_\Omega f\,. \tag{7}$$

The following theorem says that (7) is also a sufficient condition for the existence of a solution.

Theorem 1. *A solution u to the Neumann problem (2) exists if and only if f satisfies*

$$\int_\Omega f = 0\,. \tag{8}$$

Two different solutions differ by a constant.

The fact that the solution is not unique in the Neumann problem does not affect the physics of the model that is described by these equations. Take, for example, electrostatics. The electric field E^i satisfies

$$E_i = -\partial_i u, \quad \partial_i E^i = f\,, \tag{9}$$

where u is the electric potential and f the charge. If we prescribe $E^i n_i$ at the boundary we get a Neumann boundary problem for the potential u. The electric field E^i is invariant under the transformation $u \to u + c$, where c is a constant. We will see in Sect. 3 that something similar happens for the constraint equations in General Relativity.

We have seen that the Neumann problem has not a unique solution. If we include lower order terms in the operator, the Dirichlet problem will not have a unique solution either. For example, for some constants $\lambda > 0$ (the eigenvalues) the following equations have a non-trivial solutions (eigenfunctions)

$$\Delta u + \lambda u = 0 \text{ on } \Omega\,, \quad u = 0 \text{ on } \partial\Omega\,. \tag{10}$$

One of the main ideas in the theory of partial differential equations is that many relevant properties of the equations depend only on the principal part, that is on the terms with highest derivatives. The previous examples show that uniqueness does not depend only on the principal part. Motivated by the Neumann problem, we write the following the two main properties of elliptic equations

(i) The solutions space of the homogeneous problem (i.e., when we set the source f and the boundary values g equal to zero) is finite dimensional.
(ii) The solution will exist if and only if the sources satisfy a finite number of conditions.

We will see in the next sections that, under appropriate assumptions, (i)–(ii) depend only on the principal part of the equation and boundary conditions.

One example of a boundary condition that does not satisfy (i) is the following.

Example 1. Let Ω be the unit ball in $\mathbb{R}^3$ centered at the origin. Consider the following homogeneous boundary-value problem

$$\Delta u = 0 \text{ on } \Omega, \quad \partial_3 u = 0 \text{ on } \partial\Omega \, . \tag{11}$$

An explicit calculation shows that the space of solutions of this problem is *infinite dimensional* (see [27], Chap. 1, for details). In the next section we will see that this is related to the fact that the vector ∂_3 is *tangential* to the boundary at the points $x_3 = 0$, $x_1^2 + x_2^2 = 1$.

2 Second Order Elliptic Equations

Consider the following, second order, differential operator

$$Lu = \partial_i \left(a^{ij}(x)\partial_j u + b^i(x)u \right) + c^i(x)\partial_i u + d(x)u \, , \tag{12}$$

where we will assume that the coefficients are smooth functions on $\mathbb{R}^n$ and $i, j = 1, \dots n$. We have written the operator (12) in divergence form because it will be more suitable for the following calculations; since the coefficient a^{ij} and b^i are smooth, this is equivalent to the standard formula

$$Lu = a^{ij}(x)\partial_i\partial_j u + \hat{b}^j(x)\partial_j u + \hat{d}(x)u \, . \tag{13}$$

where $\hat{b}^j = \partial_j a^{ij} + b^j + c^j$ and $\hat{d} = \partial_i b^i + d$.

The principal part of the operator is given by the terms which contains only second derivatives

$$l(x, \partial) = a^{ij}(x)\partial_i\partial_j \, . \tag{14}$$

To define the symbol of L we replace in the principal part each derivative by the component of an arbitrary constant vector in $\mathbb{R}^n$

$$l(x, \xi) = a^{ij}(x)\xi_i\xi_j, \quad \xi \in \mathbb{R}^n \, . \tag{15}$$

The symbol l of L is a polynomial of order 2 in the components of ξ.

We make now the crucial assumption on the symbol. We say that the operator L is *elliptic* in $\bar{\Omega}$ if

$$l(x,\xi) \neq 0 \quad \forall x \in \bar{\Omega},\, \xi \in \mathbb{R}^n, \xi \neq 0\,. \tag{16}$$

The next important concept is the *formal adjoint* of L. The formal adjoint L^t is defined by the relation

$$\int_\Omega vLu = \int_\Omega uL^t v \tag{17}$$

for all u, v of *compact support* in Ω. In this particular case we have

$$L^t v = \partial_j \left(a^{ij}(x)\partial_i v - c^j(x)v\right) - b^i(x)\partial_i v + d(x)v \tag{18}$$

Note that $\Delta = \Delta^t$.

We have already seen in the case of the Laplacian that the solutions of the homogeneous problem play an important role; in general L and L^t are different operator and then we have two natural null spaces defined as

$$\mathcal{N}(L) = \{u : Lu = 0 \text{ on } \Omega \text{ and } u = 0 \text{ on } \partial\Omega\} \tag{19}$$

$$\mathcal{N}(L^t) = \{u : L^t u = 0 \text{ on } \Omega \text{ and } u = 0 \text{ on } \partial\Omega\}\,. \tag{20}$$

We can now formulate an existence result for the Dirichlet problem which will essentially ensure that properties (i)–(ii) are satisfied.

Theorem 2. *(i) Precisely one of the following statements holds:*
a) For each f there exists a unique solution of the boundary-value problem

$$Lu = f \text{ on } \Omega, \quad u = 0 \text{ on } \partial\Omega\,, \tag{21}$$

or else
b) $\mathcal{N}(L)$ is non-trivial.
(ii) Furthermore, should assertion b) hold, the dimension of $\mathcal{N}(L)$ is finite *and equals the dimension of $\mathcal{N}(L^t)$.*
(iii) Finally, the boundary-value problem (21) has a solution if and only if

$$\int_\Omega fv = 0 \quad \text{for all } v \in \mathcal{N}(L^t)\,. \tag{22}$$

We will consider now the analog of the Neumann problem for L. If in the integration by parts given by (17) we allow functions u and v which are not of compact support, we have to include the boundary terms; and we obtain the following relation which is called the *Green formula* for the operator L

$$\int_\Omega vLu - uL^t v = \oint_{\partial\Omega} vBu - uB^t v\,, \tag{23}$$

where the differential boundary operators are given by

$$Bu = n_j a^{ij}\partial_i u + b^i n_i u, \quad B^t v = n_j a^{ij}\partial_i v - c^i n_i v\,. \tag{24}$$

We want to solve the following problem

$$Lu = f \text{ on } \Omega, \quad Bu = 0 \text{ on } \partial\Omega \,. \tag{25}$$

As in the case of the Dirichlet problem, we define the null spaces

$$\mathcal{N}(L, B) = \{u : Lu = 0 \text{ on } \Omega \text{ and } B(u) = 0 \text{ on } \partial\Omega\} \tag{26}$$

$$\mathcal{N}(L^t, B^t) = \{u : L^t u = 0 \text{ on } \Omega \text{ and } B^t(u) = 0 \text{ on } \partial\Omega\} \,. \tag{27}$$

We have the following existence result, which looks exactly the same as the previous theorem if we replace the Dirichlet condition by the new boundary condition.

Theorem 3. *(i) Precisely one of the following statements holds:*
a) For each f there exist a unique solution of the boundary-value problem

$$Lu = f \text{ on } \Omega, \quad Bu = 0 \text{ on } \partial\Omega \,, \tag{28}$$

or else
b) $\mathcal{N}(L, B)$ is non-trivial.
(ii) Furthermore, should assertion b) hold, the dimension of $\mathcal{N}(L, B)$ is finite and equals the dimension of $\mathcal{N}(L^t, B^t)$.
(iii) Finally, the boundary-value problem (28)–(24) has a solution if and only if

$$\int_\Omega fv = 0 \quad \text{for all } v \in \mathcal{N}(L^t, B^t) \,. \tag{29}$$

We have written the boundary conditions in the form (24) in order to emphasize that they come naturally from the integration by parts. It is possible to write them in a perhaps more familiar form. Define the vector β^i by

$$\beta^i = n_j a^{ij} \,. \tag{30}$$

By the elliptic condition (16) we have $\beta^i n_i \neq 0$, that is *β^i it is never tangential to the boundary* (this excludes Example 1). In the operator L only enters the symmetric part of the matrix a^{ij}, however, we have not assumed that this matrix is symmetric in the previous theorem. If we decompose $a^{ij} = a_s^{ij} + b^{ij}$ where $a_s^{ij} = a_s^{(ij)}$ and $b^{ij} = b^{[ij]}$ is an arbitrary anti-symmetric matrix, then

$$\beta^i = n_j a_s^{ij} + \tau^i, \quad \tau^i = n_j b^{ij} \,, \tag{31}$$

where τ^i is an arbitrary tangential vector. Choosing appropriated b^i and c^i such that they do not change the operator L, we get that the function $\sigma = b^i n_i$ is also arbitrary. We conclude that the boundary condition $B(u) = 0$ is equivalent to

$$Bu = \beta^i \partial_i u + \sigma u = 0 \,, \tag{32}$$

where σ is an arbitrary function and β^i is an arbitrary non-tangential vector field on the boundary.

Let us compare Theorem 2 and 3 with the analog cases for the Laplace equation. We have now two operators L and L^t which have two different null spaces (in the case when $b^i + c^i = 0$ we have $L = L^t$ and $B = B^t$, and then only one null space). There are no statements about uniqueness or about the elements and dimension of the null spaces. We have already seen that these properties depend on the lower order terms. For the particular case of second order elliptic operators, there exists an important tool that can give uniqueness and a characterization of the null space for certain kind of lower order terms: the *maximum principle*. There exist many useful versions of the maximum principle (see for example [14]), here we mention a particular simple case, which can be generalized to other situations as we will see.

We can write the Green formula (23) in terms of a first order bilinear form $\mathbf{B}$

$$\mathbf{B}(u,v) = \oint_{\partial\Omega} vBu - \int_\Omega vLu = \oint_{\partial\Omega} uB^t v - \int_\Omega uL^t v \tag{33}$$

where

$$\mathbf{B}(u,v) = \int_\Omega (a^{ij}\partial_j u + b^i u)\partial_i v - (c^i \partial_i u + du)v \ . \tag{34}$$

From this equation we deduce that $u \in \mathcal{N}(L,B)$ if and only if $\mathbf{B}(u,v) = 0$ for all v. (One "if" is trivial; to see the other one, take test functions v which vanish at the boundary and are arbitrary in the interior). If we assume $b^i = c^i = 0$ and $d \leq 0$, then $\mathbf{B}$ is symmetric (i.e. $\mathbf{B}(u,v) = \mathbf{B}(v,u)$) and positive,

$$\mathbf{B}(u,u) \geq 0 \quad \text{for all } u \ . \tag{35}$$

Moreover, $\mathbf{B}(u,u) = 0$ if and only if u is a constant and $u = 0$ if d is not identically zero. In this case we are in a similar situation as in the Neumann problem for the Laplace equation: the only elements of the null space are the constants. More general versions of the maximum principle can be used to prove the followings refinements of Theorems 2 and 3.

Theorem 4. *Assume $d \leq 0$. Then the Dirichlet problem*

$$Lu = f \ \text{ on } \Omega, \quad u = g \ \text{ on } \partial\Omega \ , \tag{36}$$

has a unique solution for every f and g.

Theorem 5. *Assume $d \leq 0$, $\sigma \geq 0$ and not both identically zero. Let β^i a vector field such that $\beta^i n_i > 0$ on $\partial\Omega$. Then the oblique derivative problem*

$$Lu = f \ \text{ on } \Omega, \quad \beta^i \partial_i u + \sigma u = g \ \text{ on } \partial\Omega \ , \tag{37}$$

has a unique solution for every f and g.

In both theorems, the maximum principle can be used also to prove that the solution is positive if the sources and boundary values are positive.

Note that in Theorems 2 and 3 the null spaces for the operator and the adjoint have the same dimension. We will see in the next section that this will not be the case for more general operators and boundary conditions.

We conclude this section with some examples.

Example 2. The most important second order elliptic operator is the Laplacian on a Riemannian manifold. It is given by

$$Lu = \Delta_h u = h^{ij} D_i D_j u \, , \tag{38}$$

where h is a Riemannian metric ($a^{ij} = h^{ij}$) and D its corresponding covariant derivative. One important example of lower order term is given by the conformal Laplacian which appears naturally in the Einstein constraint equations

$$Lu = \Delta_h u - \frac{R}{8} u \, , \tag{39}$$

where R is the Ricci scalar of h_{ab}.

For a Riemannian metric, the principal part of the boundary condition $B(u)$ has a geometric interpretation

$$Bu = n^i D_i u, \tag{40}$$

where we use the standard convention $n^i = h^{ij} n_j$. That is, the vector n is now the unit normal vector with respect to the metric h_{ij}. This is sometimes denoted as *conormal boundary condition*.

An example of lower order boundary terms is the following

$$Bu = n^i D_i u + Hu \, , \tag{41}$$

where H is the mean curvature of the boundary Ω with respect to the metric h_{ij}. This boundary condition appears in connection to black holes (see [21] and [8]).

3 Elliptic Systems

3.1 Definition of Ellipticity

We saw in the previous section that ellipticity is a positivity condition on the symbol of the equation. In order to generalize this concept for systems of equations (this includes as particular case higher order equations) we need to define the symbol of a system. We can use the same idea as before, and define the principal part as the collection of terms which have the highest order derivatives. That is, consider the following differential operator in $\mathbb{R}^n$

$$L(u) = \sum_{|\alpha| \leq 2m} a_\alpha(x) \partial^\alpha u \, , \tag{42}$$

where α is a multi-index, and the coefficients a_α are $N \times N$ matrices. The principal part is defined as

$$l(x,\partial) = \sum_{|\alpha|=2m} a_\alpha(x)\partial^\alpha \ , \tag{43}$$

and the symbol

$$l(x,\xi) = \sum_{|\alpha|=2m} a_\alpha(x)\xi^\alpha \ . \tag{44}$$

The operator is elliptic if $\det l(x,\xi) \neq 0$ for every $x \in \bar{\Omega}$ and $\xi \neq 0$. This is the definition that appears in most text books; we will call it *classical ellipticity* (there is no general agreement on the nomenclature, in most places these systems are called just elliptic). This definition excludes many important examples; the most remarkable is perhaps the Laplace equation as a first order system (Example 3). In order to include these cases, we need to be more flexible in our definition of the principal part; in particular it is important to allow terms of different orders in it. The appearance of terms of different orders in the principal part is a particular feature of systems which is not present in higher order equations.

It will be convenient to use a more explicit notation as the one given in (43). Let $u_1, \ldots, u_N$ be functions which depend on the coordinates $x_1, \ldots, x_n$. The operator (42) can be written as follows.

$$L_{\mu\nu}(x,\partial)u^\nu(x) = f_\mu(x), \quad \nu,\mu = 1,\ldots,N \ , \tag{45}$$

where $L_{\mu\nu}$ are polynomials in $(\partial_1 \ldots, \partial_n)$ with coefficients depending on x. $[L_{\mu\nu}]$ is a $N \times N$ matrix, not necessarily symmetric. Note that N (dimension of the vectors u^ν) and n (dimension of $\mathbb{R}^n$) are in general different numbers.

Let $s_1, \ldots, s_N$, $t_1, \ldots, t_N$ be integers (some may be negative) such that

$$\deg(L_{\mu\nu}) \leq s_\mu + t_\nu \ , \tag{46}$$

where deg means the degree of the polynomial $L_{\mu\nu}$ in the derivatives ∂. The integers s_μ are attached to the equations and the t_ν to the unknowns.

We define the principal part $l_{\mu\nu}(x,\partial)$ as the terms in $L_{\mu\nu}$ which are *exactly* of order $s_\mu + t_\nu$. The symbol $l_{\mu\nu}(x,\xi)$ is obtained replacing in the principal part the derivatives by a vector ξ. We define the following polynomial in ξ.

$$l(x,\xi) = \det(l_{\mu\nu}(x,\xi)) \ . \tag{47}$$

The degree m of the systems is given by

$$m = \frac{1}{2}\deg(l(x,\xi)) \ , \tag{48}$$

where deg means degree in ξ.

The following general definition of ellipticity was introduced in [9]

Definition 1 (Douglis-Nirenberg Ellipticity). *The system* (45) *is elliptic if there exist integer weights* s_μ *and* t_ν *which satisfy* (46) *such that* $l(x,\xi) \neq 0$ *for all real* $\xi \in \mathbb{R}^n$, $\xi \neq 0$, $x \in \bar{\Omega}$, *where* $l(x,\xi)$ *is given by* (47).

For $n = 2$ we assume in addition

Definition 2 (Supplementary Condition). $l(x,\xi)$ *is of even degree* $2m$. *For every pair of linearly independent real vectors* ξ *and* ξ', *the polynomial* $l(x, \xi + \tau\xi')$ *in the complex variable* τ *has exactly* m *roots with positive imaginary parts.*

Every elliptic system in dimension $n \geq 3$ satisfies the supplementary condition (see [1]). This is no longer true for $n = 2$, as Example 5 shows. A system that is elliptic in the sense of Definition 1 and satisfies also the supplementary condition (Definition 2) will be called *properly elliptic*.

Note that the definition depends on the weights s_μ and t_ν which are not unique, a system can be elliptic for many different choices of weights. Also note that the number $2m$ is not related in general with the degree of the highest derivatives; for example for a second order system with $N = 3$ we have $m = 3$ (Example 6). The degree m is important because it gives the number of boundary conditions we have to impose in order to get a well defined elliptic problem, as we will see in the next section.

There exists an important class of elliptic operators for which the Dirichlet boundary conditions will always satisfy (i)–(ii) as we will see in the next section. These systems are given by the following definition.

Definition 3 (Strong Ellipticity). *The system is called strongly elliptic if* $s_\nu = t_\nu \geq 0$ *and there exist a constant* $\epsilon > 0$ *such that*

$$Re\,(l_{\mu\nu}(x,\xi)\eta^\mu \bar{\eta}^\nu) \geq \epsilon \eta^\mu \bar{\eta}_\mu \xi^i \xi_i \,, \tag{49}$$

for all real $\xi \in \mathbb{R}^n$ *and all complex* $\eta \in \mathbb{C}^N$.

Note that every elliptic equation (i.e., $N = 1$) is strongly elliptic. Let us discuss some examples.

Example 3 (Laplace equation as a first order system). This example was taken from [2]. Consider the Laplace equation in two dimensions

$$\partial_1^2 u + \partial_2^2 u = 0 \,. \tag{50}$$

Every equation can be written as a first order system if we introduce the derivatives of the unknown as new variables. That is, let $u_1 = \partial_1 u$ and $u_2 = \partial_2 u$. Then we have the following system ($n = 2$ and $N = 3$)

$$\partial_1 u_1 + \partial_2 u_2 = 0 \,, \tag{51}$$

$$\partial_1 u - u_1 = 0 \,, \tag{52}$$

$$\partial_2 u - u_2 = 0 \,. \tag{53}$$

In the matrix notation

$$\begin{pmatrix} 0 & \partial_1 & \partial_2 \\ \partial_1 & -1 & 0 \\ \partial_2 & 0 & -1 \end{pmatrix} \begin{pmatrix} u \\ u_1 \\ u_2 \end{pmatrix} = 0 \, . \tag{54}$$

In the classical definition, the symbol is constructed only with the terms which contain the highest order derivatives, in this case only with the terms with one derivative. Then the determinant of the symbol is

$$\begin{vmatrix} 0 & \xi_1 & \xi_2 \\ \xi_1 & 0 & 0 \\ \xi_2 & 0 & 0 \end{vmatrix} = 0 \, , \tag{55}$$

and we conclude that the system is not classically elliptic.

Take the weights $t_1 = 2$, $t_2 = t_3 = 1$, for u, u_1, u_2 and $s_1 = 0$, $s_2 = s_3 = -1$, to the first, second and third equations, respectively. Then, we have

$$\begin{vmatrix} 0 & \xi_1 & \xi_2 \\ \xi_1 & -1 & 0 \\ \xi_2 & 0 & -1 \end{vmatrix} = \xi_1^2 + \xi_2^2 \, , \tag{56}$$

and the system is elliptic with $m = 1$. Another possible choice for the weights is the following: $s_i = t_i$, with $t_1 = 1$, $t_2 = t_3 = 0$.

Since $n = 2$, we have to check also that it satisfies the supplementary condition.

$$0 = l(\xi + \tau\xi') = |\xi|^2 + 2\tau\xi^i\xi'_i + \tau^2|\xi'|^2 \tag{57}$$

implies

$$\tau_\pm = (-\cos\theta \pm i \sin\theta)|\xi'|^{-1}|\xi| \tag{58}$$

where $|\xi|^2 = \xi^i\xi_i$ and $\xi^i\xi'_i = \cos\theta|\xi||\xi'|$. That is, we have only one root with positive imaginary part.

Example 4 (Stokes system). This example was taken from [28]. The following equations appear as the stationary linearized case of the Navier-Stokes equations (see for example [34]) for the velocity u^i and the pressure p of the fluid

$$\Delta u^i - \partial^i p = 0, \quad \partial^i u_i = 0 \, . \tag{59}$$

The unknowns are u^i, p, that is $N = 4$, and we will assume $n = 3$. Then, in the matrix notation we have

$$\begin{pmatrix} \Delta & 0 & 0 & -\partial_1 \\ 0 & \Delta & 0 & -\partial_2 \\ 0 & 0 & \Delta & -\partial_3 \\ \partial_1 & \partial_2 & \partial_3 & 0 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ p \end{pmatrix} = 0 \, . \tag{60}$$

It is clear that the system is not classically elliptic. Take $t_1 = t_2 = t_3 = 2$, $t_4 = 1$ and $s_1 = s_2 = s_3 = 0$, $s_4 = -1$. Then the symbol is

$$l_{ij} = \begin{pmatrix} |\xi|^2 & 0 & 0 & -\xi_1 \\ 0 & |\xi|^2 & 0 & -\xi_2 \\ 0 & 0 & |\xi|^2 & -\xi_3 \\ \xi_1 & \xi_2 & \xi_3 & 0 \end{pmatrix} , \tag{61}$$

and we have

$$l = |\xi|^6, \quad m = 3 . \tag{62}$$

Then, the system is elliptic. Another possible choice for the weights is the following: $s_i = t_i$, with $t_1 = t_2 = t_3 = 1$ and $t_4 = 0$.

Example 5 (Cauchy-Riemann equation). We write the Cauchy-Riemann equation $L(u) = \partial_{\bar{z}} u$ in terms of the real variables $z = x + iy$,

$$L(u) = \frac{1}{2} \left(\partial_x u + i \partial_y u \right) . \tag{63}$$

We have $n = 2$, $N = 1$. The symbol $l = \xi_1 + i\xi_2$ satisfies

$$l(\xi) \neq 0 \text{ for all real } \xi \neq 0 , \tag{64}$$

hence the system is elliptic with $m = 1/2$. However, it does not satisfy the supplementary condition because $2m = 1$ is not an even number.

Example 6. Consider the following operator in $\mathbb{R}^3$, acting on three-vectors u^i,

$$L_{ij} u^j = \partial^j (\mathcal{E} u)_{ij} , \tag{65}$$

where

$$(\mathcal{E} u)_{ij} = 2\mu \partial_{(i} u_{j)} + \lambda \delta_{ij} \partial^k u_k , \tag{66}$$

and μ, λ are constants. Since in this case we have $N = n = 3$ we will use the same index notation for the index in the vectors u and in the coordinates of $\mathbb{R}^3$.

The system (65) appears in elasticity (see, for example, [20]). It also appears in General Relativity related to gauge conditions like the minimal distortion gauge (see [33]) and in the constraint equations (see [39]), usually with the choice $\mu = 1$, $\lambda = -2/3$ which makes (66) trace-free.

From (65) we deduce

$$L_{ij} u^j = \left((\mu + \lambda) \partial_i \partial_j + \mu \delta_{ij} \Delta \right) u^j , \tag{67}$$

in the matrix notation we have ($\lambda' = \mu + \lambda$)

$$L_{ij} u^j \equiv \begin{pmatrix} \lambda' \partial_1^2 + \mu \Delta & \lambda' \partial_2 \partial_1 & \lambda' \partial_1 \partial_3 \\ \lambda' \partial_1 \partial_3 & \lambda' \partial_2^2 + \mu \Delta & \lambda' \partial_1 \partial_3 \\ \lambda' \partial_1 \partial_3 & \lambda' \partial_1 \partial_3 & \lambda' \partial_3^2 + \mu \Delta \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} . \tag{68}$$

Take $s_i = t_i = 1$, the symbol is given by

$$l_{ij}(\xi) = \lambda' \xi_i \xi_j + \mu \delta_{ij} \xi^k \xi_k \ , \tag{69}$$

and

$$l = \mu^2 (2\mu + \lambda) |\xi|^6, \quad m = 3 \ . \tag{70}$$

The operator is (classically) elliptic for $\mu > 0$, $2\mu + \lambda > 0$. It is also strongly elliptic

$$l_{ij} \eta^i \bar{\eta}^j = (\lambda + \mu)(\eta^i \xi_i)(\bar{\eta}^i \xi_i) + \mu \xi^k \xi_k \eta_i \bar{\eta}^i \geq \epsilon \xi^k \xi_k \eta_i \bar{\eta}^i \ , \tag{71}$$

where $\epsilon = \min\{\mu, 2\mu + \lambda\}$.

Example 7 (Einstein constraint equations). There exist different ways of reducing the Einstein constraint equations to an elliptic system (see, for example, the recent review [4]). In the standard approach the principal part of the system is formed with the Laplace operator on a Riemannian manifold given in Example 2 and the operator that has been discussed in Example 6.

A particular interesting example is the one that has been recently used in [6] and [7] to construct new kinds of solutions. This system is not elliptic in the classical sense but it satisfies definition 1 for appropriate weights (see these references for details).

Example 8 (Witten equation). The Witten equation $\partial_{AA'} u^A = 0$ (in the spinorial notation) plays an important role in the positive mass theorem of General Relativity (cf. [37]). Solutions of this equation have been analyzed in [29] and [26].

In the matrix notation ($N = 2$, and we will assume $n = 3$) this system is given by

$$\begin{pmatrix} \partial_3 & \partial_1 + i\partial_2 \\ \partial_1 - i\partial_2 & -\partial_3 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = 0 \ . \tag{72}$$

The principal part, with weights $t_1 = t_2 = 1$, $s_1 = s_2 = 0$, is given by

$$l_{\nu\mu}(x, \xi) = \begin{pmatrix} \xi_3 & \xi_1 + i\xi_2 \\ \xi_1 - i\xi_2 & -\xi_3 \end{pmatrix}, \quad l = -(\xi_3^2 + \xi_2^2 + \xi_1^2), \quad m = 1 \ . \tag{73}$$

Then, the system is elliptic.

3.2 Definition of Elliptic Boundary Conditions

For the operator $L_{\mu\nu}$ defined in (45) we will consider boundary conditions of the following form

$$B(x, \partial)_{l\nu} u^\nu = 0, \quad l = 1, \ldots, m; \quad \nu = 1, \ldots, N; \tag{74}$$

where $B(x, \partial)_{l\nu}$ are polynomial in ∂ and m is given by (48). The order of the boundary operators, like those of the operators in (45), depends on two systems of integer weights, in this case the system t_ν already attached to the

dependent variable and a new system r_l attached to each boundary condition such that

$$\deg(B_{l\nu}) \leq r_l + t_\nu \ . \tag{75}$$

Note that r_l can be negative and also the order of the derivatives in the boundary conditions can be higher than in the operator. The principal part $b_{l\nu}$ of the boundary operator consists of the terms in $B_{l\nu}$ which are exactly of order $r_l + t_\nu$.

For a given operator L, we would like to know for which boundary operators B the solutions of the corresponding boundary-value problem will satisfy (i)–(ii). The answer to this question is given by the following definition, as we will see in the next section.

Let x_0 a point on $\partial\Omega$ and let n^i the outer normal to Ω. We consider the constant coefficient problem

$$l_{\mu\nu}(x_0, \partial)u^\nu = 0 \ , \tag{76}$$

$$b_{l\nu}(x_0, \partial)u^\nu = 0 \ , \tag{77}$$

on the half plane $(x^i - x_0^i) \cdot n_i < 0$ with boundary $(x^i - x_0^i)n_i = 0$.

Definition 4 (Complementing Condition). *We say that the complementing condition holds at x_0 if there are no nontrivial solutions of (76)–(77) of the following form:*

$$u^\nu(x) = v^\nu(\eta)e^{i\xi_j(x^j - x_0^j)} \tag{78}$$

where ξ is a any nonzero, real, vector which satisfies $\xi^i n_i = 0$, $v(\eta)$ tends to zero exponentially as $\eta \to -\infty$ and the coordinate η is defined by $\eta = (x^j - x_0^j)n_j$.

In the literature, these conditions are also called *Lopatinski-Shapiro* conditions or *covering* conditions (see [2] and [38]). Let us study some examples of boundary conditions.

Example 9 (Boundary conditions for the Laplace equation). Consider solutions of the form (78) for the Laplace equation $\Delta u = 0$. We chose coordinates in $\mathbb{R}^n$ such that $\eta = x_n$, $n^i = \delta^i_n$. Then, all the solutions of this form are given by

$$u = e^{i\xi^i x_i} e^{\pm|\xi| x_n} \ , \tag{79}$$

where ξ satisfies $\xi_n = 0$.

We consider different boundary conditions on the plane $x_n = 0$. For the Dirichlet condition $u(x_n = 0) = 0$ we get

$$u = e^{i\xi^i x_i} = 0 \ . \tag{80}$$

Since this is not possible there exists no solution of this form which satisfies the Dirichlet condition. Hence, the Dirichlet boundary condition satisfies the complementing condition.

For the Neumann condition we have $\partial_{x_n} u = 0$ at $x_n = 0$, this implies $\xi = 0$ and then the solution will not decay at infinity. Hence, the Neumann condition satisfies also the complementing condition.

Take the oblique derivative boundary condition $\beta^i \partial_i u = 0$ at $x_n = 0$. This implies

$$i(\beta_i \xi^i) = 0, \quad \beta_n |\xi| = 0 \,. \tag{81}$$

If $\beta_n \neq 0$, then $|\xi| = 0$, and the complementing condition is satisfied. This was the case studied in Sect. 2. On the other hand, if $\beta_n = 0$ (like in Example 1), then the complementing condition is not satisfied since we can always chose a vector ξ such that $\beta_i \xi^i = 0$ and we will get solutions of the form (78).

Consider now the following interesting example studied in [15]. At $x_n = 0$ we impose the boundary conditions

$$\delta u = 0 \,, \tag{82}$$

where

$$\delta u = \partial_1^2 u + \cdots + \partial_{n-1}^2 u \,, \tag{83}$$

is the Laplacian in $n-1$ dimension. From (82) we deduce the $|\xi|^2 = 0$ and then it satisfies the complementing conditions. It is also clear that $\delta^k u = 0$ where k, is an arbitrary natural number, satisfies the complementing condition. Note that in this cases the boundary operator has derivatives of higher order than the Laplace operator. On a Riemannian manifold, these conditions can be written in geometric form where δ is the intrinsic Laplacian on the boundary. Another interesting condition which also satisfies the complementing condition is the following

$$\delta u - n^i \partial_i u = 0 \,. \tag{84}$$

In this case, integrating by parts, it is easy to show that the only solutions of the homogeneous problem are the constants

$$0 = \int_\Omega u \Delta u = \oint_{\partial\Omega} u n^i \partial_i u - \int_\Omega \partial_i u \partial^i u \tag{85}$$

$$= \oint_{\partial\Omega} u \delta u - \int_\Omega \partial_i u \partial^i u \tag{86}$$

$$= - \oint_{\partial\Omega} |du|^2 - \int_\Omega \partial_i u \partial^i u \,, \tag{87}$$

where du denotes the gradient intrinsic to the boundary.

Example 10. Consider the operator discussed in Example 6. Integrating by parts we get

$$\mathbf{B}(u, v) = - \int_\Omega v^i L_{ij} u^j + \oint_{\partial\Omega} (\mathcal{E} u)_{ij} n^i v^j \tag{88}$$

where

$$\mathbf{B}(u,v) = \int_\Omega (\mathcal{E}u)^{ij}\partial_i v_j \ . \tag{89}$$

We can write the integrand in $\mathbf{B}(u,v)$ in the following form

$$(\mathcal{E}u)^{ij}\partial_i v_j = \frac{\mu}{2}\,(\mathcal{L}u)_{ij}\,(\mathcal{L}v)^{ij} + \left(\lambda + \frac{2}{3}\mu\right)\partial_k u^k \partial_l v^l \ , \tag{90}$$

where $(\mathcal{L}u)_{ij}$ is the trace-free part of $\partial_{(i}u_{j)}$, that is

$$(\mathcal{L}u)_{ij} = 2\partial_{(i}u_{j)} - \frac{2}{3}\delta_{ij}\partial_k u^k \ . \tag{91}$$

Note that $\mathbf{B}$ is symmetric, $\mathbf{B}(u,v) = \mathbf{B}(v,u)$. Using this and equation (88) we get the following Green formula:

$$\int_\Omega v^i L_{ij} u^j - u^i L_{ij} v^j = \oint_{\partial\Omega} (\mathcal{E}u)_{ij} n^i v^j - (\mathcal{E}v)_{ij} n^i u^j \ . \tag{92}$$

This is analogous to the Green formula for second order equation (23). For simplicity we have not included terms in non-divergence form in the operator. That is why we have $L = L^t$ and $B = B^t$ in (92). These extra terms can be handled in the same way as in Sect. 2.

The boundary integral in the Green formula (92) suggests that two natural boundary conditions are that of Dirichlet type

$$u^i = 0 \text{ on } \partial\Omega \ , \tag{93}$$

and the analog to the Neumann boundary condition

$$(\mathcal{E}u)_{ij} n^j = 0 \text{ on } \partial\Omega \ . \tag{94}$$

We want to prove that these boundary conditions satisfy the complementing condition. We will assume that $\mu > 0$ and $2\mu + \lambda > 0$, that is, the operator is elliptic as we have seen in Example 6. We will make also an extra assumption: $3\lambda + 2\mu \geq 0$; this implies that the integrand (90) is positive. Moreover, if $3\lambda + 2\mu > 0$,

$$B(u,u) = 0 \iff \partial_{(i}u_{j)} = 0 \ , \tag{95}$$

that is u is a Killing vector. If $3\lambda + 2\mu = 0$, then

$$B(u,u) = 0 \iff (\mathcal{L}u)_{ij} = 0 \ , \tag{96}$$

then u is a conformal Killing vector. In flat space, we know explicitly all the Killing and conformal Killing vectors. Hence, we have a characterization of the null spaces for these boundary conditions. The Killing and conformal Killing vectors are the analog of the constants for the Neumann problem for the Laplace equation.

Assume we have a solution u of the form (78). Choose Cartesian coordinates such that $\eta = x_3$. Let $L_1 = 2\pi/\xi_1$ and $L_2 = 2\pi/\xi_2$. Take as domain the infinite cubic region $x_3 \geq 0$, $0 \leq x_1 \leq L_1$, $0 \leq x_2 \leq L_2$. For this domain we use equation (88) for $u = v$. We want to prove that, on this domain, the boundary integral in (88) vanishes if we impose either (93) or (94). Using these boundary conditions we get that the integrand vanishes on the face $x_3 = 0$. The integrand also vanishes on the face $x_3 = \infty$ because the solution, by hypothesis, decays at infinity. On the other faces the integrand does not vanish. However, because of the choice of L_1 and L_2, we have that the integrand of opposite faces are identical. Then, the sum of the boundary integrals vanishes because the normal is always outwards. We conclude that $\mathbf{B}(u, u)$ should vanish. But there are no Killing or conformal Killing vectors which decay to zero at infinity. Hence the complementing condition is satisfied.

Example 11 (Boundary conditions for the Stokes system). If we multiply equations (59) by u^i and integrate by parts we get

$$0 = -\int_\Omega \partial_k u_i \partial^k u^i + \oint n^k (u^i \partial_k u_i - u_k p) \, . \tag{97}$$

Using this equation and a similar argument as in the previous example it is possible to show that the boundary conditions

$$u^i = 0 \text{ on } \partial\Omega \, , \tag{98}$$

and p unprescribed, satisfy the complementing condition (see for example [28]).

Example 12 (Dirichlet boundary conditions for strongly elliptic systems). Assume that the system is strongly elliptic; this implies $s_i = t_i = t'_i \geq 0$. The Dirichlet boundary conditions on $\partial\Omega$ are given by

$$(n^i \partial_i)^q u_j = 0, \quad q = 0, \ldots, t'_j - 1, \quad j = 1, \ldots, N; \tag{99}$$

when $t'_j = 0$, u_j goes unprescribed.

It can be proved that for every strongly elliptic system, the Dirichlet conditions (99) satisfy the complementing condition (see [2]).

In the case of equations ($N = 1$) of order $2m$ ($t_1 = m$) these conditions reduce to

$$(n^i \partial_i)^q u = 0, \quad q = 0, \ldots, m - 1 \, . \tag{100}$$

In particular, for second order equations ($m = 1$) we have $u = 0$ at the boundary. That is, we recover the familiar Dirichlet condition studied in Sects. 1 and 2. In Example 6 we have $t'_i = 1$, then $q = 0$ and the Dirichlet conditions is just

$$u^j = 0 \text{ on } \partial\Omega \, . \tag{101}$$

As an example of a higher order equation, we have the biharmonic equation

$$\Delta\Delta u = f \ , \tag{102}$$

the Dirichlet conditions are given by ($N = 1$, $m = 2$)

$$u = 0, \quad n^i \partial_i u = 0 \text{ on } \partial\Omega \ . \tag{103}$$

Example 13. In the following example (taken from [28]), the complementing condition is *not* satisfied. Consider the following problem in $\mathbb{R}^2$, where the boundary is the line $x_2 = 0$

$$\Delta\Delta u = 0 \text{ on } \Omega \quad \Delta u = \partial_2 \Delta u = 0 \text{ on } \partial\Omega \ . \tag{104}$$

For every $\xi \in \mathbb{R}$ the function

$$u(x, y) = e^{i\xi x_1 - |\xi| x_2} \tag{105}$$

is a solution.

Example 14. We have seen that for strongly elliptic systems the Dirichlet boundary conditions satisfy the complementing condition. This is not true for general elliptic systems. In the following example (discussed in [25]) we show that there are elliptic systems for which the Dirichlet problem is not well defined.

Consider the system ($N = n = 2$)

$$\begin{pmatrix} \partial_1^2 - \partial_2^2 & -2\partial_1\partial_2 \\ 2\partial_1\partial_2 & \partial_1^2 - \partial_2^2 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = 0 \ . \tag{106}$$

The symbol is given by

$$l_{ij} = \begin{pmatrix} \xi_1^2 - \xi_2^2 & -2\xi_1\xi_2 \\ 2\xi_1\xi_2 & \xi_1^2 - \xi_2^2 \end{pmatrix}, \quad l = (\xi_1^2 + \xi_2^2)^2 \ . \tag{107}$$

Then, the system is elliptic in the classical sense. This system can be written in the complex form, $z = x_1 + ix_2$, $w = u_1 + iu_2$, as

$$\partial_{\bar{z}}^2 w \equiv \frac{1}{4}(\partial_1 + i\partial_2)^2 w = 0 \ , \tag{108}$$

for which the general solution is clearly

$$w = f(z) + \bar{z} g(z) \ , \tag{109}$$

where f and g are arbitrary functions of z. We observe that all solutions of the form

$$w = f(z)(1 - z\bar{z}) \text{ on } |z| \leq 1 \ , \tag{110}$$

with arbitrary analytic f, vanish on the boundary of the unit disk. Thus, the Dirichlet boundary conditions do not characterize the solutions: there exist infinitely many solutions with identical boundary values.

Example 15 (Boundary conditions for the Witten equation). In the Witten equation studied in Example 8 we have $m = 1$, that is, we can only impose one boundary condition. Consider the following boundary condition:

$$u_1 = 0 \text{ on } \partial\Omega \tag{111}$$

and u_2 goes unprescribed. This condition has been studied in [31] as an inner boundary condition for black holes in the positive mass theorem. We want to prove that it satisfies the complementing condition. We can explicitly calculate all the solutions of the form (78) of the equations (72)

$$u^\nu = e^{\xi^i x_i} v^\nu(x_3), \quad v^\nu = A^\nu e^{|\xi| x_3} \ , \tag{112}$$

where A^ν are constants such that $A_2/A_1 = (i\xi_1 + \xi_2)|\xi|^{-1}$ and we choose coordinates such that $\eta = x_3$, $\xi_3 = 0$. There is no solution of this form that satisfies $u_1(x_3 = 0) = 0$ and then the complementing condition follows.

Example 16 (Stationary solutions of Einstein equations). In the presence of a timelike symmetry, the Einstein equations can be reduced to an elliptic system. Moreover, the inner boundary conditions satisfy the complementing condition. This result was proved in [30] and it was used to prove an existence result for the non linear problem. See also [23] for a different kind of boundary conditions for the static case.

3.3 Results

In order to present a general result for properly elliptic systems with boundary conditions that satisfy the complementing condition we need to reformulate in a more precise way properties (i)–(ii). For a given operator L and boundary operator B we consider the operator A defined as $A(u) = (L(u), B(u))$. This operator will act on appropriate Hilbert spaces $\mathcal{H}_1$ and $\mathcal{H}_2$, $A : \mathcal{H}_1 \to \mathcal{H}_2$. In analogous way to (26) we define the null space $\mathcal{N}(A)$ as

$$\mathcal{N}(A) = \{u \in \mathcal{H}_1 : A(u) = 0\} \ . \tag{113}$$

The range of A is defined by

$$\mathcal{R}(A) = \{w \in \mathcal{H}_2 : \exists u \in \mathcal{H}_1, A(u) = w\} \ , \tag{114}$$

and the complement of the range is given by

$$\mathcal{R}^\perp(A) = \{w \in \mathcal{H}_2 : (Au, w)_{\mathcal{H}_2} = 0 \text{ for all } u \in \mathcal{H}_1\} \ , \tag{115}$$

where $(\cdot, \cdot)$ denotes the Hilbert space scalar product.

We can write now properties (i)–(ii) as follows.

(i) $N(A)$ has finite dimension.
(ii) $\mathcal{R}^\perp(A)$ has finite dimension.

An operator which satisfies (i)–(ii) is called a *Fredholm* operator. (We have assumed that A is bounded, otherwise in (ii) we need to impose that $\mathcal{R}(A)$ is closed, see [18] and [16]).

We have the following general result. (We only sketch the statement, for details and proofs see [17, 16] and also [38].)

Theorem 6. *If the system L is properly elliptic in $\bar{\Omega}$ and the boundary conditions satisfy the complementing condition for every point of $\partial\Omega$, then the operator $A(u) = (L(u), B(u))$ is Fredholm.*

We have seen that the dimension of $\mathcal{N}$ (and hence uniqueness) is not invariant if we add lower order terms to the operator. One of the consequence of Theorem 6 is the existence of an invariant for elliptic problems: the Fredholm index. This number is defined as

$$I = \dim \mathcal{N}(A) - \dim \mathcal{R}^{\perp}(A) \; . \tag{116}$$

It can be proved that the index I is stable under perturbation, in particular it does not depend on the lower order terms.

In Sect. 2 we have used the Green formula to construct the formal adjoint operator L^t and its corresponding boundary operator B^t. In this case we can define $A^t(u) = (L^t(u), B^t(u))$, and it can be proved[1] that $N(A^t) = \mathcal{R}^{\perp}(A)$. That is, the boundary-value problems considered in Theorem 2 and 3 have $I = 0$. In fact these theorems also show that the index does not depend on the lower order terms in this particular case.

Boundary conditions which come from a Green formula are called *normal boundary conditions*. The advantage of them is that we have a characterization of $\mathcal{R}^{\perp}(A)$ through the formal adjoint problem, and then we can in principle compute the conditions that the sources should satisfy in order to have a solution. General results for normal boundary conditions for higher order elliptic equations can be found in [19, 32]. Since these boundary conditions come from an integration by parts, the order of the boundary operators will be always less than that of the operator itself. We have seen that this is not necessarily the case for general elliptic boundary conditions that satisfy the complementing condition. For the general case, we will not have a characterization of $\mathcal{R}^{\perp}(A)$.

We have seen that the Dirichlet boundary conditions satisfy the complementing condition for strongly elliptic systems. Using this fact, general existence results for the Dirichlet problem can be proved (see [24]). Moreover, it can be shown that the index is always zero in this case.

Finally, we want to present an existence result for the operator considered in Example 6 that can be deduced from the general Theorem 6 (see [36]).

[1] It is important to note that for any bounded (or unbounded with dense range) operator A we can define the Hilbert adjoint A'. This is not related, in general, with the formal adjoint A^t. However, when we have a Green formula, it is possible to prove that in fact $A^t = A'$ (see Theorem 8.4 of [19, 5] and also [35]).

In this case, we have a Green formula and then we have normal boundary conditions. The following two theorems are analogous to Theorem 2 and 3.

Theorem 7. *Let L_{ij} be given by (65) with $\mu \geq 0$, $2\mu + \lambda \geq 0$, $2\mu + 3\lambda \geq 0$. Then, for every smooth f^j and g^i, there exists a unique, smooth, solution u^i of the Dirichlet problem*

$$L_{ij}u^i = f^j \text{ on } \Omega, \quad u^i = g^i \text{ on } \partial\Omega \, . \tag{117}$$

We have seen that all solutions of the homogeneous problem satisfy $(\mathcal{E}v)_{ij} = 0$, that is v is a Killing or a conformal Killing vector. Uniqueness in this theorem follows because there exists no Killing or conformal Killing vector which vanishes at the boundary.

Theorem 8. *Let L_{ij} be given by (65) with $\mu \geq 0$, $2\mu + \lambda \geq 0$, $2\mu + 3\lambda \geq 0$. Consider the boundary-value problem*

$$L_{ij}u^i = f^j \text{ on } \Omega, \quad (\mathcal{E}u)_{ij}u^i = 0 \text{ on } \partial\Omega \, . \tag{118}$$

This problem has a solution if and only if

$$\int_\Omega f_i v^i = 0 \quad \text{for all } v^i \text{ such that } \mathcal{E}v_{ij} = 0 \, . \tag{119}$$

If u_1 and u_2 are two different solutions, then the difference $v = u_1 - u_2$ satisfies $(\mathcal{E}v)_{ij} = 0$.

In the case of the Einstein constraint equations, the previous theorems can be used to prove existence of solutions of the momentum constraint (see [39]). In this case the physical quantity is the second fundamental form K_{ij} which is given by

$$K_{ij} = Q_{ij} - (\mathcal{E}u)_{ij} \, , \tag{120}$$

where Q is an (essentially) arbitrary tensor. Then, as in the case of the Neumann problem for the Laplace equation, the lack of uniqueness in Theorem 8 will not affect K_{ij}.

4 Final Comments

In order to check if a system of equations is elliptic, we should first prove that the principal part of the operator satisfies definitions 1 and 2. If the system is non linear, we should consider the corresponding linearized problem. Then we should prove that the boundary operators satisfy the complementing condition (definition 4). This can be complicated. There exist equivalent formulations of this condition (see for example [38]), some of them can be more suitable for specific problems. It is also important to know if the boundary conditions come from a Green formula (normal boundary conditions). The

Green formula can be used to prove that the complementing condition holds (as we have seen in the examples). Moreover, for the case of higher order equations there exist general results that can be used (see [19]).

We have only discussed linear elliptic systems. In the non linear case, there are no general existence results like Theorem 6. For non linear second order equations a good reference is [14] and for non linear systems [13] and [12]. A related issue that was not discussed here is regularity. We have assumed that the functions and the boundary are smooth. Regularity properties are crucial for non linear systems, see [14, 13] and [12] and references there.

In Sect. 2 we have followed [11]. Good references for this section are also [10, 14] and [22]. For Sect. 3, an introductory book is [28]; more advanced material can be found in [35] and [3]; for a complete discussion see [16, 15] and [38].

Acknowledgements

I would like to thank the organizers of the "319th WE-Heraeus-Seminar", J. Frauendiener, D. Giulini and V. Perlick, for the invitation. Part of the material for this contribution was presented in two lectures at the school on "Structure and dynamics of compact objects" which took place in the Albert Einstein Institut on 20-25/09, 2004; as a part of the SFB/TR7 project. I would also like to thank the organizers S. Boutloukos, J. Frauendiener and S. Husa for the invitation. Finally, thanks to F. Beyer and D. Giulini for their careful reading of the manuscript.

This work has been supported by the Sonderforschungsbereich SFB/TR 7 of the Deutsche Forschungsgemeinschaft.

References

1. S. Agmon, A. Douglis, L. Nirenberg: Estimates near the boundary for solutions of elliptic partial differential equations satisfying general boundary conditions. I. Comm. Pure App. Math. **12**, 623–727 (1959) 125
2. S. Agmon, A. Douglis, L. Nirenberg: Estimates near the boundary for solutions of elliptic partial differential equations satisfying general boundary conditions. II. Comm. Pure App. Math. **17**, 35–92 (1964) 125, 129, 132
3. M.S. Agranovich: Elliptic boundary problems. In: *Partial Differential Equations, IX*, volume 79 of *Encyclopaedia Math. Sci.* (Springer, Berlin 1997) pp 1–144 137
4. R. Bartnik, J. Isenberg: The constraint equations. In: *The Einstein Equations and Large Scale Behavior of Gravitational Fields*, ed by P.T. Chruściel and H. Friedrich (Birhäuser, Basel 2004) pp 1–38. `gr-qc/0405092` 128
5. F.E. Browder: Estimates and existence theorems for elliptic boundary value problems. Proc. Nat. Acad. Sci. U.S.A. **45**, 365–372 (1959) 135

6. P.T. Chruściel, E. Delay: On mapping properties of the general relativistic constraints operator in weighted function spaces, with applications. (2003) `gr-qc/0301073` 128
7. J. Corvino, R. M. Schoen: On the asymptotics for the vacuum Einstein constraint equations. (2003) `gr-qc/0301071` 128
8. S. Dain: Trapped surfaces as boundaries for the constraint equations. Class. Quantum. Grav. **21**(2), 555–573 (2004) `gr-qc/0308009` 123
9. A. Douglis, L. Nirenberg: Interior estimates for elliptic systems of partial differential equations. Comm. Pure App. Math. **8**, 503–538 (1955) 124
10. L.C. Evans: *Partial Differential Equations*, volume 19 of *Graduate Studies in Mathematics* (American Mathematical Society, Providence, Rhode Island 1998) 137
11. G.B. Folland: *Introduction to Partial Differential Equation* (Princeton University Press, Princeton, New Jersey 1995) 137
12. M. Giaquinta: *Multiple Integrals in the Calculus of Variations and Nonlinear Elliptic Systems*, volume 105 of *Annals of Mathematics Studies*. (Princeton University Press, Princeton, New Jersey 1983) 137
13. M. Giaquinta: *Introduction to Regularity Theory for Nonlinear Elliptic Systems*. Lectures in Mathematics, ETH Zürich (Birkhäuser, Basel 1993) 137
14. D. Gilbarg, N.S. Trudinger: *Elliptic Partial Differential Equations of Second Order* (Springer, Berlin 1983) 122, 137
15. L. Hörmander: *Linear Partial Differential Operators*, volume 116 of *Grundlehren der Mathematischen Wissenschaften* (Academic Press, New York 1963) 130, 137
16. L. Hörmander: *The Analysis of Linear Partial Differential Operators. III*, volume 274 of *Grundlehren der Mathematischen Wissenschaften* (Springer, Berlin 1985) 135, 137
17. L. Hörmander: *The Analysis of Linear Partial Differential Operators. IV*, volume 275 of *Grundlehren der Mathematischen Wissenschaften* (Springer, Berlin 1985) 135
18. T. Kato: *Perturbation Theory for Linear Operators*, volume 132 of *Grundlehren der Mathematischen Wissenschaften*. (Springer, Berlin 1980) 135
19. J.L. Lions, E. Magenes: *Non-Homogeneous Boundary Value Problems and Applications.*, volume 181 of *Grundlehren der Mathematischen Wissenschaften* (Springer, New York 1972) 135, 137
20. J.E. Marsden, T. J. Hughes: *Mathematical Foundations of Elasticity*, Prentice-Hall Civil Engineering and Engineering Mechanics Series (Prentice Hall, Englewood Cliffs, New Jersey 1983) 127
21. D. Maxwell: Solutions of the Einstein constraint equations with apparent horizon boundary. (2003) `gr-qc/0307117` 123
22. R.C. McOwen: *Partial Differential Equations* (Prentice Hall, Englewood Cliffs, New Jersey 1996) 137
23. P. Miao: On existence of static metric extensions in general relativity. Commun. Math. Phys. **241(1)**, 27–46 (2003) 134
24. L. Nirenberg: Remarks on strongly elliptic partial differential equations. Comm. Pure App. Math. **8**, 649–675 (1955) 135
25. L. Nirenberg: Estimates and existence of solutions of elliptic equations. Comm. Pure App. Math. **9**. 509–529 (1956) 133
26. T. Parker, C. H. Taubes: On Witten's proof of the positive energy theorem. Comm. Math. Phys. **84**(2), 223–238 (1982) 128

27. P.R. Popivanov, D.K. Palagachev: *The degenerate oblique derivative problem for elliptic and parabolic equations*, volume 93 of *Mathematical Research* (Akademie-Verlag, Berlin 1997) 119
28. M. Renardy, R. C. Rogers: *An Introduction to Partial Differential Equations*, 2nd edn, volume 13 of *Texts in Applied Mathematics* (Springer, New York 2004) 126, 132, 133, 137
29. O. Reula: Existence theorem for solutions of Witten's equation and nonnegativity of total mass. J. Math. Phys. **23**(5), 810–814 (1982) 128
30. O. Reula: On existence and behaviour of asymptotically flat solutions to the stationary Einstein equations. Commun. Math. Phys. **122**, 615–624 (1989) 134
31. O. Reula, K. P. Tod: Positivity of the Bondi energy. J. Math. Phys. **25(4)**, 1004–1008 (1984) 134
32. M. Schechter: General boundary-value problems for elliptic partial differential equations. Comm. Pure Appl. Math. **12**, 457–486 (1959) 135
33. L. Smarr, J. York: Radiation gauge in general relativity. Phys. Rev. D **17**, 1945–1956 (1978) 127
34. H. Sohr: *The Navier-Stokes equations*. Birkhäuser Advanced Texts (Birkhäuser, Basel 2001) 126
35. M.E. Taylor: *Partial Differential Equations. I*, volume 115 of *Applied Mathematical Sciences* (Springer, New York 1996) 135, 137
36. J.L. Thompson: Some existence theorems for the traction boundary-value problem of linearized elastostatics. Arch. Rational Mech. Anal. **32**, 369–399 (1969) 135
37. E. Witten: A new proof of the positive energy theorem. Comm. Math. Phys. **80(3)**, 381–402 (1981) 128
38. J.T. Wloka, B. Rowley, B. Lawruk: *Boundary Value Problems for Elliptic Systems* (Cambridge University Press, Cambridge 1995) 129, 135, 136, 137
39. J.W. York: Conformally invariant orthogonal decomposition of symmetric tensor on Riemannian manifolds and the initial-value problem of general relativity. J. Math. Phys. **14**(4), 456–464 (1973) 127, 136

Mathematical Properties of Cosmological Models with Accelerated Expansion

Alan D. Rendall

Max-Planck-Institut für Gravitationsphysik, Am Mühlenberg 1, 14476 Golm, Germany
rendall@aei.mpg.de

Abstract. An introduction to solutions of the Einstein equations defining cosmological models with accelerated expansion is given. Connections between mathematical and physical issues are explored. Theorems which have been proved for solutions with positive cosmological constant or nonlinear scalar fields are reviewed. Some remarks are made on more exotic models such as the Chaplygin gas, tachyons and k-essence.

1 Introduction

Recent cosmological observations indicate that the expansion of the universe is accelerating and this has led to a great deal of theoretical activity. Models of accelerated cosmological expansion also raise a variety of interesting mathematical questions. The purpose of the following is to first give a pedagogical introduction to this subject suitable for the mathematically inclined reader and then to present an overview of some of the mathematical results which have been obtained up to now and the many challenges which remain.

The simplest class of cosmological models consists of those with the highest symmetry, i.e. those which are homogeneous and isotropic. The underlying spacetimes are the FLRW (Friedmann-Lemaître–Robertson–Walker) models. A further simplification can be achieved by assuming that the metric of the slices of constant time is flat. The spacetime metric can be written in the form:

$$-dt^2 + a^2(t)(dx^2 + dy^2 + dz^2) \tag{1}$$

for a suitable scale factor $a(t)$. These are the models most frequently used in the literature due both to their simplicity and the fact that spatially flat FLRW models appear to give a good description of our universe.

The physical interpretation of $a(t)$ is that if two typical galaxies are a distance $D(t)$ apart at time t then $D(t_2)/D(t_1) = a(t_2)/a(t_1)$ for any times t_1 and t_2. The statement that the universe is expanding corresponds to the condition that the time derivative $\dot{a}$ is positive. Accelerated expansion means that the second derivative $\ddot{a}$ is positive.

The function $a(t)$ should be determined by the field equations for gravity and in the following we always take the Einstein equations for this purpose.

A.D. Rendall: *Mathematical Properties of Cosmological Models with Accelerated Expansion*, Lect. Notes Phys. **692**, 141–155 (2006)
www.springerlink.com

There are two choices to be made. The one concerns the cosmological constant Λ. The other concerns the description of the matter content of spacetime. This means choosing the variables which describe the matter, the equations of motion these are to satisfy and the definition of the energy-momentum tensor as a function of the matter fields and the spacetime metric. Under the assumption of FLRW symmetry this will lead to an evolution equation for $a(t)$. The easiest way to produce models with accelerated expansion is to choose a positive cosmological constant ($\Lambda > 0$). A more sophisticated alternative is to choose $\Lambda = 0$ but to include a suitable nonlinear scalar field among the matter fields.

The rest of this article is structured as follows. It starts with a brief introduction to some physical ideas relevant to accelerated cosmological expansion. Then mathematical theorems about spacetimes with positive cosmological constant motivated by the physics are described. After that these results are compared with the original physical motivation. Once the case of a positive cosmological constant has been described it is discussed why it might be good to replace the cosmological constant by a nonlinear scalar field and what changes when that is done. Finally, some future research directions involving more general models for cosmic acceleration are indicated. In particular, comments are made on the Chaplygin gas and the tachyon condensate.

2 Physical Background

Accelerated expansion plays a role in cosmology in two different regimes. The first is the very early universe while the second is the period between the decoupling of the microwave background radiation and the present. Accelerated expansion in the early universe is associated with the name inflation which was introduced by Guth [10]. The paper [10] was extremely important in popularizing the concept of inflation. A first-hand account of the historical development of the idea of inflation can be found in [11] where there is information on related earlier work of other authors such as Starobinsky ([11], p. 229).

One of the attractive features of inflation is that it is claimed to solve certain 'problems' in cosmology. It is justified to ask in which sense these can really be considered as problems but these philosophical questions will not be entered into in the following. Among these issues are

– homogeneity and isotropy,
– flatness problem,
– horizon problem.

The first issue is that, after averaging on a suitable scale, our universe is homogeneous and isotropic. There are two basically different kinds of possible reason for this. One is that it was always homogeneous and isotropic. This

possibility is perceived by many as unsatisfactory. The alternative is that the universe was originally anisotropic and inhomogeneous and that some dynamical mechanism later made it homogeneous and isotropic. In the second explanation this mechanism must be found. The second issue is that it appears that the curvature of space on cosmological scales is very small today. Within a standard FLRW model this implies that it was even smaller at decoupling. It is often perceived that this smallness requires an explanation. The third issue is that the temperature of the microwave background is essentially the same at points such that there would have been no time to send a signal to both from some common point since the big bang in a standard Friedmann model (without accelerated expansion). Can this be explained? Inflation has something to say about all three issues, as will be shown later.

Accelerated cosmological expansion at the present epoque is a relatively recent discovery, dating from the late 1990's. There is now very strong observational evidence, which continues to accumulate, that the velocity of recession of distant galaxies is accelerating. On the theoretical side this phenomenon is associated with the names dark energy and quintessence. The latter term was introduced by Caldwell, Dave and Steinhardt [5]. There are a number of different lines of evidence for cosmic acceleration at times after decoupling which include

- supernovae of type Ia,
- microwave background fluctuations,
- gravitational lensing,
- galaxy clustering.

Here only the supernova data will be discussed. A supernova of type Ia is an exploding star which is bright enough to be visible at cosmological distances. The characteristics of an event of this type which can in principle be observed are the red shift, the light curve (observed brightness as a function of time) and the spectrum. In recent years it has become possible to observe these data in practice for a useful sample of objects. The light curve and spectrum provide the information needed to identify a supernova as being of type Ia. The advantage of this is that type Ia supernovae have universal properties which allow their intrinsic brightness to be determined. In a first approximation, all of these objects have the same intrinsic brightness at the maximum of their light curves. The number of objects of this type observed so far is just over 150. The projected space mission SNAP (Supernova Acceleration Probe) is planned to observe about 2000 per year. The way in which data can be compared with theoretical models will be outlined in Sect. 6.

3 Mathematical Developments

It has been known for a long time that spacetimes with a positive cosmological constant have a tendency to isotropize at late times, a circumstance

associated with the name 'cosmic no hair theorem'. In [23] Starobinsky wrote down formal expansions for the late-time behaviour of spacetimes with positive cosmological constant. He studied the case where the matter is described by a perfect fluid with linear equation of state $p = (\gamma - 1)\rho$, where γ is a constant belonging to the interval $[1, 2)$. He also discussed the vacuum case which, as it turns out, gives the leading order terms in the expansion of the geometry for the case with fluid as well. In a certain sense the solutions all look like the de Sitter solution at late times. This will be made more precise below.

It will be convenient in the following to write the de Sitter solution in the form

$$-dt^2 + e^{2Ht}(dx^2 + dy^2 + dz^2) \tag{2}$$

where $H = \sqrt{\Lambda/3}$. These coordinates only cover half of de Sitter space but this is no disadvantage in the following where the subject of interest is the limit $t \to \infty$. The expansions of [23] are expressed in terms of Gauss coordinates. In other words $g_{00} = -1$ and $g_{0a} = 0$ where Latin indices are spatial indices. In the vacuum case the expansion of the spatial metric is

$$g_{ab}(t, x) = e^{2Ht}(g^0_{ab}(x) + g^2_{ab}(x)e^{-2Ht} + g^3_{ab}(x)e^{-3Ht} + \cdots) \, . \tag{3}$$

The fact that the coefficient of e^{Ht} vanishes is a result of the analysis. Putting $g^0_{ab} = \delta_{ab}$ and setting all the other coefficients to zero gives the de Sitter solution.

In [23] it is not specified how this infinite series is to be interpreted mathematically but it is natural to interpret it as a formal series. This means that there is no assertion that the series converges or even that it is asymptotic. Recall that a series as above is called asymptotic if for each positive integer M there is a positive constant C_M such that

$$\left| g_{ab} - \sum_0^M g^m_{ab} e^{(2-m)Ht} \right| \leq C_M e^{(1-M)Ht} \, . \tag{4}$$

In other words, the sum of any finite truncation of the series differs from the quantity to which it is asymptotic by a remainder of order equal to the next term beyond the truncation. A convergent series is asymptotic but not necessarily conversely. At this point in the discussion it is not even claimed that the above series is asymptotic. It is just a formal expression which solves the Einstein equations in the sense that if we substitute it into the Einstein equations and manipulate the infinite series according to rather obvious rules all terms cancel.

In [20] a theorem was proved concerning the above formal series. To formulate it, let A_{ab} be a three-dimensional Riemannian metric and B_{ab} a symmetric tensor which is transverse traceless with respect to A_{ab}. This means that $A^{ab}B_{ab} = 0$ and $\nabla^a B_{ab} = 0$ where the covariant derivative is that associated to the metric A_{ab}. Given A_{ab} and B_{ab} of this form which are smooth

(C^∞) there exists a unique series of the above form satisfying the vacuum Einstein equations with $\Lambda > 0$ with smooth coefficients g^m_{ab} which satisfies the conditions $g^0_{ab} = A_{ab}$ and $g^3_{ab} = B_{ab}$. Notice that on the basis of function counting these solutions are as general as the general solution of the vacuum Einstein equations. For the general solution can be specified by giving the induced metric and the second fundamental form on a spacelike hypersurface and these must satisfy one scalar and one vector equation. Thus in both cases we have the same type of data and the same number of constraints which they have to satisfy. In the present case the constraints are simpler than in the ordinary Cauchy problem. In [20] a corresponding theorem was also proved for the case of the Einstein equations coupled to a perfect fluid with a linear equation of state. The most difficult part of the proof is to show that the Einstein constraint equations are satisfied as a consequence of the 'constraints at infinity', i.e. the tranverse traceless nature of B_{ab} with respect to A_{ab}.

It is desirable to extend the above results about formal power series and function counting to show that there exists a large class of solutions which have asymptotic expansions of the above form and that these are general in the sense that they include all solutions arising from a non-empty open set of initial data on a Cauchy surface. One place to look for such an open set is as an open neighbourhood of standard data for the de Sitter solution on a hypersurface $t = \text{const}$. In the vacuum case a result of this kind was proved in [20] using results of Friedrich [6, 7] on the stability of de Sitter space. The corresponding result with a perfect fluid, which is what would be desirable for cosmology, is still open. The proofs in the vacuum case use the conformal field equations. The results can be extended in some cases to conformally invariant matter fields but for other matter fields, including most fluids, it is not at all clear that the method could work. If the metric is conformally rescaled as in the conformal method either the rescaled metric or the conformal factor will for most fluids be non-smooth, involving non-integral powers of the time coordinate.

Another possible direction in which the existing results could be extended is to other spacetime dimensions. In the context of formal power series of the vacuum Einstein equations with $\Lambda > 0$ this has been done in [20]. The result is the series:

$$g_{ab} = e^{2Ht}\left(g^0_{ab} + \sum_{m=1}^{\infty}\sum_{l=0}^{L_m}(g_{ab})_{m-2,l}t^l e^{-mHt}\right) \tag{5}$$

where $H = \sqrt{2\Lambda/n(n-1)}$ in spatial dimension n. For each m the quantity L_m is a finite integer. The terms with $l > 0$ will be refererred to as 'logarithmic terms' since t is logarithmic in the expansion parameter e^{Ht}. Again it is possible to prescribe two quantities A_{ab} and B_{ab} which this time have to satisfy an inhomogeneous version of the transverse traceless condition in general. The inhomogeneity is determined by A_{ab}. The prescribed coefficients

are $g^0_{ab} = A_{ab}$ and $(g_{ab})_{n-2,0} = B_{ab}$. In general logarithmic terms are required to get a consistent formal expansion. They can only be avoided if n is odd, $n = 2$ or A_{ab} satisfies some strong restrictions.

At the present time the results on formal asymptotic expansions for higher dimensional vacuum spacetimes have not been extended to existence theorems for all solutions corresponding to a non-empty open set of initial data on a regular Cauchy surface. It has, however, been proved that there exists a very large class of solutions of the Einstein equations with asymptotic expansions as above. Tensors A_{ab} and B_{ab} satisfying the constraints at infinity can be prescribed arbitrarily under the assumption that they are analytic (C^ω). This was proved in [20] using Fuchsian techniques. The generality of the solutions is judged using function counting. These results can probably be extended to fluids with linear equation of state in $3 + 1$ dimensions but this has not been worked out.

The above results require no symmetry assumptions. Under the assumption of spatial homogeneity much more is known. A theorem of Wald [27] shows that for spacetimes of Bianchi types I-VIII with positive cosmological constant and matter satisfying the dominant and strong energy conditions solutions which exist globally in the future have certain asymptotic properties as $t \to \infty$. This implies that the asymptotics of these spacetimes have some of the properties which follow from the asymptotic expansions discussed above. To go further the matter model must be specified. For matter described by the Vlasov equation global existence and more refined asymptotics have been proved by Lee [16]. When the matter model is a perfect fluid with linear equation of state similar results have been proved in [21]. These results confirm many of the features expected from the formal asymptotic expansions. There is also a class of highly symmetric inhomogeneous spacetimes with $\Lambda > 0$ for which global existence and asymptotic properties has been proved for large initial data. These are solutions of the Einstein-Vlasov system with plane or hyperbolic symmetry [24, 25].

4 Mathematics and Physics Compared

In all the classes of spacetimes with a positive cosmological constant which expand forever the available mathematical results all indicate isotropization at late times. To see the reason for this, introduce the second fundamental form of the hypersurfaces $t = \text{const.}$, which in Gauss coordinates is given by $k_{ab} = -(1/2)\partial_t g_{ab}$. It turns out that the tracefree part of k_{ab} becomes negligible in comparison with its trace $\text{tr}k$, which is the mean curvature. Equivalently each eigenvalue of the second fundamental form divided by the mean curvature tends to $1/3$ as $t \to \infty$. In the FLRW models these values are exactly equal to $1/3$. In the terminology more common in general relativity the ratio of shear to expansion tends to zero. This is the meaning of isotropization.

At first sight it seems that the spacetime does not become homogeneous at late times, since the coefficient g^0_{ab} of the leading term in the expansion is not homogeneous. There is, however, a more subtle sense in which it does become homogeneous. Globally in space there is certainly no uniform convergence to a homogeneous metric. This is also the case for spacetime regions of constant coordinate size in the Gaussian coordinates which have been used. On a spatial region of fixed physical size, however, things look different. A region of this kind has a coordinate size which goes to zero exponentially. Since any Riemannian metric can be approximated arbitrarily well by a flat metric on a sufficiently small region it follows that on a region of fixed physical size the metric converges uniformly and exponentially to the de Sitter metric. In this sense the spacetime does become homogeneous.

Consider next the flatness problem. If the metric has an asymptotic expansion of the form given in the last section then it can be computed directly that the scalar curvature of the spatial metric converges to zero exponentially as $t \to \infty$ and this is what we want to solve the flatness problem. In fact even more can be said. The curvature invariants $R_{ab}R^{ab}$ and $R_{abcd}R^{abcd}$ associated with the three-dimensional metric also decay exponentially. Thus it is not just the scalar curvature which decays; the entire curvature of the spatial metric decays just as fast. It should be noted that although the results of [24] and [25] give a lot of information on the spacetimes to which they are applicable, they are apparently not strong enough to give curvature decay.

It is not so easy to address the horizon problem by a simple and precise mathematical statement. What can be said is the following. A positive cosmological constant leads to solutions of the Einstein equations which look like de Sitter space on a long time interval and a long time interval in de Sitter space does not suffer from the horizon problem.

5 Scalar Fields

As already mentioned in the introduction, an alternative to a positive cosmological constant as a mechanism for producing solutions of the Einstein equations with accelerated expansion is a suitable nonlinear scalar field. Consider a minimally coupled scalar field in a spacetime with vanishing cosmological constant. The energy-momentum tensor of the scalar field is of the form

$$T_{\alpha\beta} = \nabla_\alpha\phi\nabla_\beta\phi - \left[\frac{1}{2}\nabla^\gamma\phi\nabla_\gamma\phi + V(\phi)\right] g_{\alpha\beta} \tag{6}$$

where V is a smooth non-negative function, the potential. To see the connection with a cosmological constant, consider the spatially homogeneous case. Then the energy density is given by $\rho = T_{00} = \dot{\phi}^2/2 + V(\phi)$ while the pressure is given by $p = T_{11} = \dot{\phi}^2/2 - V(\phi)$. There are now different possible regimes. If the kinetic energy is much larger than the potential energy on

a certain time interval then on that interval the energy density is approximately equal to the pressure. Thus in a certain loose sense the matter can be approximated by a stiff fluid, which satisfies $p = \rho$. If the kinetic and potential energies are approximately equal on a certain time interval then the pressure is approximately zero there. On that interval the matter can be approximated by dust, which satisfies $p = 0$. Finally, if the potential energy is much larger than the kinetic energy then the pressure is approximately equal to minus the energy density. It is the third case which is related to a cosmological constant. If we think of the cosmological constant as a matter field whose energy-momentum tensor is proportional to the metric then this fictitious matter satisfies $p = -\rho$. In particular, the pressure is negative and comparable in size to the energy density and this is what gives rise to accelerated expansion.

The nature of the dynamics with a nonlinear scalar field depends crucially on the potential V. A useful intuitive picture for guessing what happens with a given potential is the 'rolling' picture. In any spatially homogeneous spacetime the equation of motion for the scalar field is

$$\ddot{\phi} - (\mathrm{tr}k)\dot{\phi} + V'(\phi) = 0 \tag{7}$$

This is similar to the equation of motion of a ball which rolls on the graph of the function V with variable friction determined by the mean curvature $\mathrm{tr}k$. Physical intuition then suggests that the ball should roll down the slope and settle down in a local minimum of the potential. It turns out that accelerated expansion eventually stops if the minimum value of the potential is zero and that for that reason the case of a strictly positive minimum is mathematically more tractable.

Depending on how the acceleration of the universe varies with time it may or may not be consistent with the simplest model where there is a positive cosmological constant and any other matter present satisfies the strong energy condition and so cannot by itself cause acceleration. If the obervations are not consistent with acceleration caused only by a cosmological constant then the next simplest possibility is the nonlinear scalar field. Whether a cosmological constant is enough to explain the observations does not yet seem to be settled although there is some work on the problem in the literature. (See e.g. [1].)

There have been many suggestions for the form of the potential V in the context of inflation or quintessence but there is no clear winner at the moment. If there is a scalar field causing cosmological expansion then we do not know what it is. In these circumstances it makes sense to study the properties of large classes of potentials. In [21] the case of a potential with a strictly positive minimum was discussed. For spacetimes containing a scalar field of this type and ordinary matter satisfying the dominant and strong energy conditions it was shown that there are rather direct generalizations of Wald's theorem [27] and the results of [16].

More specifically, it can be shown under weak assumptions that if the potential is bounded below by a positive constant then $V'(\phi)$ tends to zero

as $t \to \infty$. Either ϕ converges to a finite value which is a critical point of V or ϕ tends to plus or minus infinity. If ϕ converges to a finite value and if the corresponding critical point of V is a non-degenerate local minimum ϕ_1 then the solution has asymptotics like that in Wald's theorem, with $V(\phi_1)$ playing the role of an effective cosmological constant. The mean curvature tr$k$ converges to a constant $-3H_1$.

In [21] the statement was made that when the potential has a non-degenerate positive minimum a solution for which the scalar field converges to this minimum has no oscillations. This is misleading and should be replaced by the statement that the deviation of the scalar field from the point where the potential attains its minimum and the modulus of $\dot{\phi}$ decay exponentially as $t \to \infty$. This implies in particular that $\dot{\phi}$ is absolutely integrable. The equation for $u = (\phi, \dot{\phi})$ can be written in the form $\dot{u} = Au + R(t)u$ where A is a constant matrix and $R(t)$ is a matrix-valued function which decays exponentially. If $\beta > 9H_1^2/4$ where $\beta = V''(\phi_1)$ then the eigenvalues of A are not real. For a generic solution there is an oscillation modulating the leading order exponential decay of the scalar field.

A natural next step is to look at potentials which are strictly positive but which are allowed to go to zero at infinity. The best-studied case is that of power-law inflation. Analogues of Wald's theorem for this case were obtained in [14] and extended in [17]. The potential is of the form $V = V_0 e^{-\kappa\lambda\phi}$ for a positive constant λ. Here κ is a constant which in geometrical units ($G = c = 1$) satisfies $\kappa^2 = 8\pi$. Accelerated expansion at late times is obtained if $\lambda < \sqrt{2}$. If λ is greater than $\sqrt{2}$ then the expansion is decelerated at late times. In the accelerated case the scale factor behaves like a power of t greater than one at late times. When $\lambda > \sqrt{2}$ there are exact FLRW models where the scale factor is proportional to a power of t which is less than one [12].

For inhomogeneous models there is just one interesting result. In [19] formal series expansions for spacetimes with power-law inflation and matter content given by a scalar field alone were written down. It would be interesting to extend the results of [20] for a cosmological constant to this case. The formal expansions are more complicated since they can include powers which are any linear combination with integer coefficients of one and λ. This is similar to the case of a perfect fluid where integer linear combinations of one and γ occur. Note that there is at present no analogue of the results of [6] and [7] known for the case of power-law inflation. It would also be interesting to extend the results of [25] to the case of a nonlinear scalar field. A first step in this direction is a local existence theorem for solutions of the Einstein equations coupled to the Vlasov equation and a linear scalar field which was obtained in [26].

If the potential is zero somewhere the dynamical behaviour becomes more complicated. This is what happens in chaotic inflation. The model case is that of a massive linear scalar field. There is accelerated expansion on some finite time interval but it eventually stops, a process known as reheating. After

this the scalar field behaves like dust. At late times $\dot{\phi}$ does not decay faster than t^{-1} and so is not absolutely integrable. These conclusions are based on heuristic arguments [3].

6 Relations Between Perfect Fluids and Scalar Fields

A type of matter model frequently used to produce accelerated expansion is a perfect fluid which violates the strong energy condition. The equation of state $p = f(\rho)$ satisfies $\rho + 3p < 0$. In the simplest case of a linear equation of state $p = (\gamma - 1)\rho$ this corresponds to choosing $\gamma < 2/3$. Unfortunately $\gamma < 1$ means that $dp/d\rho < 0$ and so the speed of sound becomes imaginary. As has been argued in [8] this suggests that for inhomogenous solutions the initial value problem is ill-posed. The case of homogeneous spacetimes should be thought of as a simple and important special case of the problem without symmetry and if the model makes no mathematical sense without symmetry it is suspect.

A solution to this difficulty is the observation that there is a certain equivalence between a perfect fluid and a scalar field and that the scalar field defines a model which is well-posed without symmetry restrictions. Consider first the case of a linear equation of state with $0 < \gamma < 2/3$ and no other matter fields. Suppose that a spatially flat FLRW solution is given for a fluid with this equation of state. We look for a potential such that the corresponding nonlinear scalar field can reproduce the fluid solution. Using the equation of state gives the relation $\dot{\phi}^2 = \frac{2\gamma}{2-\gamma}V$. Differentiating this with respect to time gives an equation relating $\ddot{\phi}$ and $V'(\phi)$. All terms in this equation have a common factor $\dot{\phi}$. Because $p \neq \rho$ it follows that $\dot{\phi} \neq 0$ and this factor can be cancelled. It follows that $\ddot{\phi} = \frac{\gamma}{2-\gamma}V'(\phi)$. The Hamiltonian constraint implies that $\mathrm{tr}k = -\sqrt{\frac{48\pi V}{(2-\gamma)^2}}$. Putting all this information into the equation of motion for the scalar field gives the equation $V' = -\sqrt{24\pi\gamma}V$. Solving this equation shows that $V = V_0 e^{-\sqrt{24\pi\gamma}\phi}$. Thus the only kind of potential which can work is the one we have already seen for power-law inflation, with $\lambda = \sqrt{3\gamma}$. The range of values of λ which occurs is exactly that which we already saw. It can be shown that this potential really does reproduce the fluid solution. To see this, notice that the initial data which must be chosen for the scalar field are uniquely determined by the data for the fluid. The quantities p and ρ defined from this scalar field satisfy the Euler equations since the energy-momentum tensor of the scalar field is divergence-free. Hence they agree with the fluid density and pressure everywhere. Note that this procedure does not extend to models which are homogeneous but not isotropic.

The above analysis can be generalized to other equations of state. Consider again the case of a perfect fluid and no other matter fields. Some general

assumptions will be made on the equation of state to make a smooth and complete discussion possible. It should, however, be noted that the considerations which follow may be usefully applied in more general situations. Here it is assumed that $dp/d\rho < C_1 < 1$ for a constant C_1 and $p/\rho > C_2 > -1$ for a constant C_2. Note that for any nonlinear scalar field $|p/\rho| \leq 1$. For a general equation of state the relation

$$\frac{1}{2}\dot{\phi}^2 - V(\phi) = f\left(\frac{1}{2}\dot{\phi}^2 + V(\phi)\right) \tag{8}$$

must be analysed. This can be rewritten in the form $F(\frac{1}{2}\dot{\phi}^2, V(\phi)) = 0$. Suppose that we have one solution of this equation. The implicit function theorem gives the existence of a function g which satisfies $F(g(V), V) = 0$ for V close to its value in the original solution. This is because the partial derivative of F with respect to its first argument is non-zero. The function g satisfies the relation

$$g'(V) = \frac{1 + f'(g(V) + V)}{1 - f'(g(V) + V)} \,. \tag{9}$$

As a consequence the derivative of the locally defined function g remains bounded on its domain of definition and g can be extended to a longer interval provided it does not tend to zero at the endpoint of the interval. If g tended to zero then this would imply that $p/\rho \to -1$, in contradiction to what has been assumed concerning the equation of state. It follows that the relation (8) can be inverted globally to give $\dot{\phi}^2 = 2g(V)$. Following the same steps as in the case of a linear equation of state gives the equation

$$V'(\phi) = -\frac{1}{1 + g'(V)}\sqrt{48\pi g(V)(g(V) + V)} \,. \tag{10}$$

An exotic fluid model for accelerated cosmological expansion is the Chaplygin gas [13] with equation of state $p = -A/\rho$ for a positive constant A. It satisfies $dp/d\rho > 0$ but violates the dominant energy condition for $\rho < A$, since in that case $p/\rho < -1$. It is ruled out by the assumptions made above. It turns out, however that there are cosmological models with this equation of state where $\rho > A$ everywhere so that this difficulty is avoided. The calculations as above can be done for the Chaplygin gas assuming the inequality $\rho > A$. The result is surprisingly simple. The potential is given by $V(\phi) = \frac{1}{2}\sqrt{A}(\cosh\sqrt{24\pi}\phi + \frac{1}{\cosh\sqrt{24\pi}\phi})$. Thus a potential is obtained which has a strictly positive lower bound and it satisfies the hypotheses of the theorems of [21]. Thus for the scalar field corresponding to the Chaplygin gas detailed information is available about late time asymptotics. Unfortunately, because of the fact that the transformation to the scalar field picture is not globally defined, it is not possible to immediately deduce full information on the late-time dynamics for the Chaplygin gas. It should also be remembered

that the correspondence with a scalar field does not apply to solutions of the Einstein equations with a Chaplygin gas which are homogeneous but not isotropic.

Sometimes it is desirable to parametrize the degrees of freedom in a cosmological model with fluid in a way which is different from that using the equation of state. An important example of this is the machinery required to compare supernova observations with theoretical models. This will now be sketched. If we know both the apparent and intrinsic brightness of a source then we can compute its distance. (Technically, what can be computed is the so-called luminosity distance.) Consider a spatially flat FLRW model. Then the redshift z of an object is given in terms of the scale factor by $1 + z = a(t_o)/a(t_e)$, where t_o is the time at which the light from the object is observed and t_e the time at which it is emitted. In a model of this kind the luminosity distance can be computed to be $D_L = ra(t_o)(1 + z)$, where r is the spatial separation between the wordlines of observer and emitter as measured in standard coordinates. Let $H = -\mathrm{tr}k/3$, the Hubble parameter. If the luminosity distance and Hubble parameter are expressed in terms of redshift then the following relation results [22]:

$$H(z) = \left[\frac{d}{dz}\frac{D_L(z)}{1+z}\right]^{-1} . \tag{11}$$

Supernova data provides points on the curve $D_L(z)$ and the equation (11) in principle then determines $H(z)$.

Let us ignore complications due to having only discrete data and suppose we know the function $D_L(z)$ exactly. It will now be shown how the scale factor $a(t)$ can be reconstructed. Firstly, $H(z)$ can be computed using (11). An elementary computation shows that $dt/dz = -[H(z)(1 + z)]^{-1}$. Integrating this gives t as a function of z and inverting this gives z as a function of t. Thus $H(t)$ can be determined. Integrating once more gives $a(t)$. In practise, in order to distinguish between different theoretical models, an ansatz is made for $H(z)$ containing some parameters and a best fit analysis of the data is carried out to obtain values for these parameters.

7 Tachyons and Phantom Fields

The ordinary scalar field we have considered up to now can be derived from a Lagrangian with density $-\nabla_\alpha\phi\nabla^\alpha\phi - V(\phi)$. Recently dark energy models have been considered where the Lagrangian density is a more general nonlinear function $p(\nabla_\alpha\phi\nabla^\alpha\phi, \phi)$. This is known as k-essence [2]. A great advantage of the ordinary nonlinear scalar field is that it is guaranteed to have well-behaved dynamics in the full inhomogeneous case. The Cauchy problem is always well-posed. (This is even true if the potential is allowed to be negative.) In contrast, k-essence models need not have a well-posed

local Cauchy problem. The equation of motion of the scalar field need not be hyperbolic. An additional complication is that since the equations are in general quasilinear rather than semilinear (the propagation speed of waves depends on the solution), the scalar field may develop shocks. In this case there is an additional source of singularities supplementing the familiar ones in general relativity. A useful discussion of some of these points, and the question of which energy conditions are satisfied by k-essence models, can be found in [9]. The models which violate the dominant energy condition are called phantom or ghost models.

An interesting example is given by the case where the function p is given by $V(\phi)\sqrt{1+\nabla_\alpha\phi\nabla^\alpha\phi}$, which is known as the tachyon field or tachyon condensate. Note that although the word 'tachyon' originally denoted a particle which travels faster than light, the tachyon field considered here has no superluminal propagation. All characteristics of the equation lie inside the light cone. The tachyon condensate corresponds to an effective field theory for a large collection of tachyons. Consider now the special case where $V(\phi)$ is identically one. Then provided the gradient of ϕ is timelike this model is equivalent to a special case of the Chaplygin gas. To see this it suffices to define the four-velocity of the fluid by

$$u^\mu = \frac{\nabla^\mu\phi}{\sqrt{-\nabla_\alpha\phi\nabla^\alpha\phi}}\,. \tag{12}$$

This velocity field is irrotational. The equation for a Chaplygin gas in four-dimensional Minkowski space also describes a timelike hypersurface of zero mean curvature (a membrane) in five-dimensional Minkowski space. Questions of global existence for these equations have been studied in [15].

8 Closing Remarks

This paper gives a general introduction to the subject of cosmological models with accelerated expansion, taking a mathematical point of view. After some basic concepts have been introduced, the relevant physical background on inflation and quintessence is outlined. After this, various existing mathematical results in the case of a positive cosmological constant are presented. They are then confronted with the physical motivation. The exposition continues with a review of results in the case where the cosmological constant is replaced by a nonlinear scalar field. Some interesting open problems are mentioned. There are close relations between models with scalar fields and models with perfect fluids whose equation of state is more or less exotic. Some of these connections are explained. Following this it is explained how scalar fields defined by Lagrangians which are non-linear in the first derivatives give rise to models (known as k-essence) which various connections to both more conventional scalar fields (with are linear in derivatives) and perfect fluids.

At this moment new observations on cosmic acceleration are stimulating a vigorous model-building activity. One aspect of this is that if string theory is a theory of everything then it should, in particular, be able to explain dark energy. It is thus natural that string theory should be one of the main sources of new models. There are many models which are not touched on at all in this paper, in particular those coming from brane-world scenarios [18] or loop quantum cosmology [4]. We have taken a conservative strategy which covers some of the models which are easier to understand mathematically. Even with these limitations we could only treat a few aspects of the subject. A useful task for mathematical relativity is to establish clear definitions of the various models and to identify interesting dynamical issues concerning the solutions. Another task is to systematize the web of relations which exists relating different models and to determine which of them are (in an appropriate sense) really different. Apart from its pedagogical aspects this paper is intended to be a step towards meeting these challenges.

Acknowledgements

I thank Yann Brenier for useful discussions. I gratefully acknowledge the support of the Erwin Schrödinger Institute, Vienna, where part of the research for this paper was carried out.

References

1. U. Alam, V. Sahni, T.D. Saini, A.A. Starobinsky: Is there supernova evidence for dark energy metamorphosis? Mon. Not. Roy. Astron. Soc. **354**, 275–291 (2004) 148
2. C. Armendariz-Picon, V. Mukhanov, P. Steinhardt: Essentials of k-essence. Phys. Rev. D **63**, 103510 (2001) 152
3. V.A. Belinskii, L.P. Grishchuk, Ya. B. Zeldovich, I.M. Khalatnikov: Inflationary stages in cosmological models with a scalar field. Sov. Phys. JETP **62**, 195–203 (1986) 150
4. M. Bojowald: Loop quantum cosmology: recent progress (2004) `gr-qc/0402053` 154
5. R.R. Caldwell, R. Dave, P.J. Steinhardt: Cosmological imprint of an energy component with general equation of state. Phys. Rev. Lett. **80**, 1582–1585 (1998) 143
6. H. Friedrich: Existence and structure of past asymptotically simple solutions of Einstein's field equations with positive cosmological constant. J. Geom. Phys. **3**, 101–117 (1986) 145, 149
7. H. Friedrich: On the global existence and asymptotic behaviour of solutions to the Einstein-Yang-Mills equations. J. Diff. Geom. **34**, 275–345 (1991) 145, 149
8. H. Friedrich, A.D. Rendall: The Cauchy problem for the Einstein equations. In : *Einstein's Field Equations and their Physical Implications*, ed by B. G. Schmidt (Springer, Berlin 2000) 150

9. G.W. Gibbons: Phantom matter and the cosmological constant. (2003) hep-th/0302199 153
10. A.H. Guth: The inflationary universe: a possible solution to the horizon and flatness problems. Phys. Rev. D **23**, 347–356 (1981) 142
11. A.H. Guth: *The Inflationary Universe* (Perseus Books, Reading 1997) 142
12. J.J. Halliwell: Scalar fields in cosmology with an exponential potential. Phys. Lett. B **185**, 341–344 (1987) 149
13. A. Kamenshchik, U. Moschella, V. Pasquier: An alternative to quintessence. Phys. Lett. B **511**, 265–268 (2001) 151
14. Y. Kitada, K. Maeda: Cosmic no-hair theorem in homogeneous spacetimes I. Bianchi models. Class. Quantum Grav. **10**, 703–734 (1993) 149
15. H. Lindblad: A remark on global existence for small initial data of the minimal surface equation in Minkowskian space time. Proc. Amer. Math. Soc. **132**, 1095–1102 (2004) 153
16. H. Lee: Asymptotic behaviour of the Einstein-Vlasov system with a positive cosmological constant. Math. Proc. Camb. Phil. Soc. **137**, 495–509 (2004) 146, 148
17. H. Lee: The Einstein-Vlasov system with a scalar field (2004) gr-qc/0404007 149
18. R. Maartens: *Brane-World Gravity*. Living Reviews in Relativity **7** (2004), 7. http://www.livingreviews.org/lrr-2004-7 154
19. V. Müller, H.-J. Schmidt, A.A. Starobinsky: Power-law inflation as an attractor solution for inhomogeneous cosmological models. Class. Quantum Grav. **7**, 1163–1168 (1990) 149
20. A.D. Rendall: Asymptotics of solutions of the Einstein equations with positive cosmological constant. Ann. H. Poincaré **5**, 1041–1064 (2004) 144, 145, 146, 149
21. A.D. Rendall: Accelerated cosmological expansion due to a scalar field whose potential has a positive lower bound. Class. Quantum Grav. **21**, 2445–2454 (2004) 146, 148, 149, 151
22. V. Sahni, T.D. Saini, A.A. Starobinsky, U. Alam: Statefinder – a new geometrical diagnostic of dark energy. JETP Lett. **77**, 201–206 (2003) 152
23. A.A. Starobinsky: Isotropization of arbitrary cosmological expansion given an effective cosmological constant. JETP Lett. **37**, 66–69 (1983) 144
24. S.B. Tchapnda, N. Noutchegueme: The surface-symmetric Einstein-Vlasov system with cosmological constant (2003) gr-qc/0304098 146, 147
25. S.B. Tchapnda, A.D. Rendall: Global existence and asymptotic behaviour in the future for the Einstein-Vlasov system with positive cosmological constant. Class. Quantum Grav. **20**, 3037–3049 (2003) 146, 147, 149
26. D. Tegankong, N. Noutchegueme, A.D. Rendall: Local existence and continuation criteria for solutions of the Einstein-Vlasov-scalar field system with surface symmetry (2004) gr-qc/0405039 149
27. R.M. Wald: Asymptotic behaviour of homogeneous cosmological models with cosmological constant. Phys. Rev. D **28**, 2118–2120 (1983) 146, 148

The Poincaré Structure and the Centre-of-Mass of Asymptotically Flat Spacetimes

László B. Szabados

Research Institute for Particle and Nuclear Physics, Hungarian Academy of Sciences, 1525 Budapest 114, P. O. Box 49, Hungary
lbszab@rmki.kfki.hu

Abstract. The asymptotic symmetries and the conserved quantities of asymptotically flat spacetimes are investigated by extending the canonical analysis of vacuum general relativity of Beig and Ó Murchadha. It is shown that the algebra of asymptotic Killing symmetries, defined with respect to a given foliation of the spacetime, depends on the fall-off rate of the metric. It is only the Lorentz Lie algebra for slow fall-off, but it is the Poincaré algebra for $1/r$ or faster fall-off. The energy-momentum and (relativistic) angular momentum are defined by the value of the Beig–Ó Murchadha Hamiltonian with lapse and shift corresponding to asymptotic Killing vectors. While this energy-momentum and spatial angular momentum reproduce the familiar ADM energy-momentum and Regge–Teitelboim angular momentum, respectively, the centre-of-mass deviates from that of Beig and Ó Murchadha. The new centre-of-mass is conserved, and, together with the spatial angular momentum, form an anti-symmetric Lorentz tensor which transforms just in the correct way under asymptotic Poincaré transformations of the asymptotically Cartesian coordinate system.

1 Introduction

Conserved quantities in various areas of physics play distinguished role, because they reduce the number of equations of motion to solve. In particular, in mechanical systems with only a few degrees of freedom the conserved quantities can (and e.g. in the Kepler problem do) specify the whole dynamics. It is true that in (not completely integrable) field theories they do not, but they can be used to parameterize the solutions of the field equations. In many cases they provide an essential characterization of the states of the physical system. For example, in Newtonian astrophysics the classification of stars is based on their total mass and the total angular momentum with respect to their own centre-of-mass, which classification is essential in the sense that even the qualitative feature of the history of the stars depends critically on the value of these parameters.

Apart from cosmology, both in general relativity and in non-gravitational physics primarily we are interested in localized systems. These systems are

L.B. Szabados: *The Poincaré Structure and the Centre-of-Mass of Asymptotically Flat Spacetimes*, Lect. Notes Phys. **692**, 157–184 (2006)
www.springerlink.com

modeled by appropriately decaying fields near infinity, whenever physical quantities, like total energy-momentum and angular momentum, can be associated with the whole system. However, as is well known, we should make a distinction between null infinity and spatial infinity. If we are interested e.g. in radiative problems then null infinity and physical quantities defined there will have significance. The familiar physical quantities are not conserved in general, rather they change in time characterizing the main aspects of the dynamics, telling us e.g. how much energy is carried away by radiation. (For a recent review see e.g. [1].) On the other hand, if we are interested only in the structure of the theory, e.g. to understand the gauge freedom or the genuine conserved quantities in the theory, then we usually consider decaying at spatial infinity. (For a possible, viable unification of the null and spatial infinities and the connection between these two, see for example [2].) One of the most natural frameworks in which these quantities are introduced is based on the Hamiltonian [3, 4, 5]. Several remarkable statements have been proven on their properties [6, 7, 8], among which the most important is probably the positive energy theorem [9, 10] and its extensions.

However, the recent investigations of the energy-momentum and (relativistic) angular momentum at the *quasi-local* level raised the question of whether or not these are the "ultimate" expressions that any reasonable quasi-local expression should reproduce at spatial infinity. (For a general discussion of these questions see e.g. [11], and for a recent, potentially promising particular expression for the centre-of-mass see [12, 13].) In fact, a systematic reexamination of these classical results showed that although the energy-momentum and the spatial angular momentum expressions seem to be the "ultimate" ones, the centre-of-mass should probably be completed by an additional (time dependent) term [14].

The main goal of the present contribution is to give a more detailed discussion of those issues of [14] that were not spelled out in detail. In particular, we extend and refine the analysis and the results of Beig and Ó Murchadha [5] on the structure of asymptotically flat spacetimes, and, especially, on the relativistic centre-of-mass. The novelty of the present approach is that we define the total energy-momentum and relativistic angular momentum as the value of the boundary term in the Beig–Ó Murchadha Hamiltonian using the $3+1$ parts of *asymptotic Killing vectors as the lapse and the shift*. This makes it possible to find the correct *explicit* time dependence of the Hamiltonian, yielding the familiar energy-momentum and spatial angular momentum, but the centre-of-mass deviates from the Beig–Ó Murchadha expression by a term which is the linear momentum times the coordinate time. We will see that the angular momentum 4-tensor built from the spatial angular momentum and the corrected centre-of-mass has much better transformation and conservation properties than the previous expressions.

Many questions in connection with the gravitational energy-momentum and (relativistic) angular momentum can be formulated even in connection

with the matter fields in Minkowski spacetime too, and it could be interesting and useful to compare the gravitational and the non-gravitational cases. Thus in Sect. 2 we discuss matter fields in Minkowski spacetime, and then, only in Sect. 3, we consider general asymptotically flat spacetimes. That section is devoted to the evolution equations and the boundary conditions. In Sect. 4 we recall the main points of the analysis and results of Beig and Ó Murchadha [5], and we formulate our questions.

The key objects in the present investigations are the asymptotic Killing vectors. These will be introduced and discussed in Sect. 5. In Sect. 6 we return to the discussion of the Beig–Ó Murchadha Hamiltonian, but, instead of the original time independent lapses and shifts, we use the lapse and shift parts of the asymptotic Killing vectors. Finally, in Sect. 7, we define the total energy-momentum and relativistic angular momentum and discuss their transformation and conservation properties. We summarize the main results in Sect. 8.

Although we aimed at giving a logically complete treatise, several important issues, e.g. the discussion of the background (in) dependence of the physical quantities, had to be left out. These can be found in [14]. We consider metrics with faster than $1/r$ fall-off as well. If the conditions of the positive energy theorem are satisfied then these fast fall-off metrics correspond only to flat spacetime configurations. However, in our investigations the 3-space Σ is not assumed to be complete, and its inner boundaries are not assumed to be marginally trapped surfaces. Hence the positive energy theorem does not imply flatness for fast fall-off. Thus it might be worth considering the fast fall-off case as well.

We use the abstract index formalism, and only the underlined and boldface indices take numerical values. The signature of the spacetime metric is -2, and the Riemann and Ricci tensors and the curvature scalar e.g. of the spacetime covariant derivative ∇_e will be defined by $-{}^4R^a{}_{bcd}X^bY^cZ^d := \nabla_Y(\nabla_Z X^a) - \nabla_Z(\nabla_Y X^a) - \nabla_{[Y,Z]}X^a$, ${}^4R_{ab} := {}^4R^c{}_{acb}$ and ${}^4R := {}^4R_{ab}g^{ab}$, respectively. Thus Einstein's equations take the form ${}^4G_{ab} := {}^4R_{ab} - \frac{1}{2}{}^4R g_{ab} = -\Lambda g_{ab} - \kappa T_{ab}$, and we use the units in which $c = 1$.

2 Symmetries and Conserved Quantities in Minkowski Spacetime

2.1 The Killing Fields of the Minkowski Spacetime

It is well known that the Killing vectors of the Minkowski spacetime form a ten dimensional Lie algebra $\mathcal{K}$, which contains a four dimensional commutative ideal $\mathcal{T}$, and the quotient $\mathcal{K}/\mathcal{T}$ is isomorphic to $so(1,3)$. The elements of $\mathcal{T}$ are the *constant* vector fields, called the translations, which inherit a natural Lorentzian metric from g_{ab}. If a point o of the Minkowski spacetime is fixed, then the quotient $\mathcal{K}/\mathcal{T}$ can be identified as the Lie algebra of those

Killing fields that are vanishing at o: They are the rotation-boost Killing vectors. Thus while the ideal of the constant vector fields is canonically determined by the geometric structure of the spacetime, the quotient $\mathcal{K}/\mathcal{T}$ can be realized by Killing fields only if the 'origin' o has been specified.

If an orthonormal basis $\{E^a_{\underline{a}}\}$, $\underline{a} = 0, \dots, 3$, of *constant* vector fields and the "origin" o have been chosen, then the familiar Cartesian coordinate system $\{x^{\underline{a}}\}$ is fixed by $E^a_{\underline{a}} = (\partial/\partial x^{\underline{a}})^a$ and $x^{\underline{a}}(o) = 0$. (Underlined Roman indices from the beginning of the alphabet are concrete *spacetime name* indices.) Thus this is not only a coordinate system in the sense of differential topology, but it has a metrical content as well. Obviously, if we change the vector basis by a Lorentz transformation, $E^a_{\underline{a}} \mapsto E^a_{\underline{b}} \Lambda^{\underline{b}}{}_{\underline{a}}$, and the origin o is shifted to a new point, then the Cartesian coordinates change according to the Poincaré transformation: $x^{\underline{a}} \mapsto x^{\underline{b}} \Lambda_{\underline{b}}{}^{\underline{a}} + C^{\underline{a}}$, where $\Lambda^{\underline{c}}{}_{\underline{b}} \Lambda_{\underline{c}}{}^{\underline{a}} = \delta^{\underline{a}}_{\underline{b}}$, and $C^{\underline{a}} \in \mathbb{R}^4$ characterizes the shift of the origin.

If the basis vector $E^a_{\mathbf{0}}$ is future pointing and timelike, then we usually write the Cartesian coordinates as $x^{\underline{a}} = (t, x^{\mathbf{i}})$, $\mathbf{i} = 1, 2, 3$. Thus the boldface Roman indices from the middle of the alphabet are concrete *spatial name* indices. In a fixed Cartesian coordinate system the general form of a Killing 1-form, given both in its covariant and its $3+1$ forms, is

$$\begin{aligned} K_a &= T_{\underline{a}} \nabla_a x^{\underline{a}} + M_{\underline{a}\,\underline{b}} \left(x^{\underline{a}} \nabla_a x^{\underline{b}} - x^{\underline{b}} \nabla_a x^{\underline{a}} \right) \\ &= \left(2x^{\mathbf{k}} M_{\mathbf{ki}} + T_{\mathbf{i}} - 2t M_{\mathbf{i0}} \right) \nabla_a x^{\mathbf{i}} + \left(2x^{\mathbf{k}} M_{\mathbf{k0}} + T_{\mathbf{0}} \right) \nabla_a t. \end{aligned} \tag{1}$$

This is a linear combination of the independent translation and rotation-boost Killing 1-forms, $K^{\underline{a}}_a := \nabla_a x^{\underline{a}}$ and $K^{\underline{a}\,\underline{b}}_a := x^{\underline{a}} \nabla_a x^{\underline{b}} - x^{\underline{b}} \nabla_a x^{\underline{a}}$, respectively, by constant coefficients $T_{\underline{a}}$ and $M_{\underline{a}\,\underline{b}} = -M_{\underline{b}\,\underline{a}}$. $T_{\mathbf{0}}$, $T_{\mathbf{i}}$, $M_{\mathbf{ij}}$ and $M_{\mathbf{i0}}$ are the components of the time and space translations, and the rotation and boost parts of K_a, respectively, in the coordinates $\{x^{\underline{a}}\}$. Note that the spatial components (in the $3+1$ form) of the boost Killing 1-forms depend linearly not only on the spatial coordinates, but on the Cartesian time coordinate as well.

2.2 Quasi-Local Energy-Momentum and Angular Momentum

Let Σ be any smooth, compact, spacelike hypersurface with smooth boundary $\mathcal{S} := \partial\Sigma$. If t^a is its future directed unit timelike normal, $\mathrm{d}\Sigma$ is the induced volume element on Σ and T^{ab} is the energy-momentum tensor of the matter fields, then we can form the flux integrals

$$\mathtt{Q}^m_\Sigma [K^a] := \int_\Sigma K_a T^{ab} t_b \mathrm{d}\Sigma \tag{2}$$

for any vector field K^a. If, however, K^a is a Killing vector, then $\mathtt{Q}^m_\Sigma[K^a]$ is conserved in the sense that if Σ' is another compact spacelike hypersurface with the same boundary $\mathcal{S}$, then the flux integrals defined on Σ and Σ'

coincide. In particular, if $D(\Sigma)$ is the domain of dependence of Σ and ξ^a is a "general time axis" compatible with a foliation Σ_t of $D(\Sigma)$ (in the sense that the Lie dragging of one leaf of the foliation along the integral curves of ξ^a with a given parameter value is another leaf), then the Lie derivative of $\mathtt{Q}^m_{\Sigma_t}[K^a]$ along ξ^a is vanishing provided that K^a is a Killing field. Therefore, for Killing vectors K^a the flux integral (2) is in fact associated with the closed spacelike 2-surface $\mathcal{S}$: $\mathtt{Q}^m_\Sigma[K^a] = \mathtt{Q}^m_\mathcal{S}[K^a]$. Note that the lapse function N of the foliation Σ_t is vanishing on $\mathcal{S}$, and the shift vector N^a is tangent to $\mathcal{S}$ on $\mathcal{S}$. The "general time axis" ξ^a need not be timelike or related to the symmetry generators K^a in any way.

Since $\mathtt{Q}^m_\mathcal{S}[K^a]$ is linear in K_a, by (1) in a fixed Cartesian coordinate system it has the structure $\mathtt{Q}^m_\mathcal{S}[K^a] = T_{\underline{a}}\,\mathtt{P}^{\underline{a}} + M_{\underline{a}\,\underline{b}}\,\mathtt{J}^{\underline{a}\,\underline{b}}$. The coefficients of the parameters $T_{\underline{a}}$ and $M_{\underline{a}\,\underline{b}}$ define the quasi-local energy-momentum and (relativistic) angular momentum of the matter fields, respectively, associated with the closed spacelike 2-surface $\mathcal{S}$. If $\mu := T^{ab}t_a t_b$ and $j^a := P^a_b T^{bc} t_c$ are the energy-density and the momentum density of the matter fields seen by the observer t^a, where $P^a_b := \delta^a_b - t^a t_b$ is the orthogonal projection to Σ, then these quasi-local quantities can be given explicitly in terms of the independent translation and rotation-boost Killing vectors as

$$\mathtt{P}^{\underline{a}} = \int_\Sigma K^{\underline{a}}_a\,(\mu t^a + j^a)\,\mathrm{d}\Sigma, \qquad \mathtt{J}^{\underline{a}\,\underline{b}} = \int_\Sigma K^{\underline{a}\,\underline{b}}_a\,(\mu t^a + j^a)\,\mathrm{d}\Sigma\,. \tag{3}$$

(For a more detailed discussion of these concepts see e.g. [11].) These integrals depend on the choice for the Cartesian coordinate system, but it is easy to see that under the Poincaré transformation $x^{\underline{a}} \mapsto x^{\underline{b}}\,\Lambda_{\underline{b}}{}^{\underline{a}} + C^{\underline{a}}$ of the coordinates $\mathtt{P}^{\underline{a}}$ and $\mathtt{J}^{\underline{a}\,\underline{b}}$ transform just in the expected correct way: $\mathtt{P}^{\underline{a}} \mapsto \mathtt{P}^{\underline{b}}\,\Lambda_{\underline{b}}{}^{\underline{a}}$ and $\mathtt{J}^{\underline{a}\,\underline{b}} \mapsto \mathtt{J}^{\underline{c}\,\underline{d}}\,\Lambda_{\underline{c}}{}^{\underline{a}}\,\Lambda_{\underline{d}}{}^{\underline{b}} + \mathtt{P}^{\underline{c}}\,(C^{\underline{a}}\,\Lambda_{\underline{c}}{}^{\underline{b}} - C^{\underline{b}}\,\Lambda_{\underline{c}}{}^{\underline{a}})$. Note that, as a consequence of the special linear time dependence of the boost Killing fields in (1), the centre-of-mass part $\mathtt{J}^{\mathbf{i}0}$ of the angular momentum also depends on the Cartesian time coordinate. Without this time dependence it would not be conserved and would not have the correct transformation properties.

2.3 Total Energy-Momentum and Angular Momentum

The flux integral (2) can be defined even if Σ is not compact, e.g. if it extends to spatial infinity of the Minkowski spacetime, provided the integral exists. To ensure the finiteness of this integral, i.e. to have finite total energy-momentum and (relativistic) angular momentum given by (3), certain boundary conditions must be imposed on the energy-density μ and momentum density j^a on Σ. Such a boundary condition e.g. on a $t = \text{const}$ hyperplane in the Cartesian coordinates $x^{\underline{a}} = (t, x^{\mathbf{i}})$ might be the *fall-off conditions*

$$\mu = \frac{1}{r^4}\mu^{(4)}\Big(t, \frac{x^{\mathbf{k}}}{r}\Big) + o\big(r^{-4}\big)\,, \tag{4}$$

$$j^{\mathbf{i}} = \frac{1}{r^4}j^{\mathbf{i}(4)}\Big(t, \frac{x^{\mathbf{k}}}{r}\Big) + o\big(r^{-4}\big)\,, \tag{5}$$

for some functions $\mu^{(4)}$ and $j^{\mathbf{i}(4)}$, where $r^2 := \delta_{\mathbf{ij}}x^{\mathbf{i}}x^{\mathbf{j}}$, the square of the radial distance in the hyperplane Σ, and $o(r^{-k})$ denotes a function $f(r)$ for which $\lim_{r\to\infty}(r^k f(r)) = 0$. $o(r^{-0})$ denotes logarithmic fall-off and $o(r^{+0})$ logarithmic divergence. We will use $O(r^{-k})$ to denote a function $f(r)$ for which the limit $\lim_{r\to\infty}(r^k f(r))$ exists. These fall-off conditions ensure the finiteness of the total energy-momentum, but the angular momentum is still diverging logarithmically. Thus to have finite total angular momentum as well, stronger or additional conditions must be imposed. One apparently natural condition could be to require slightly faster than $1/r^4$ fall-off in (4) and (5). Since, however, the typical fall-off rate of the energy and momentum densities of the electromagnetic field is $1/r^4$, by a faster fall-off condition we would exclude the electromagnetic field from our investigations. Thus we retain the $1/r^4$ fall-off, and seek for additional conditions.

Evaluating the total angular momentum expression with the energy and momentum densities satisfying (4) and (5), one arrives at the additional necessary and sufficient conditions

$$\oint_{\mathcal{S}} v^{[\mathbf{i}}j^{\mathbf{j}]}\,\mathrm{d}\mathcal{S}_1 = o\big(r^{-4}\big)\,, \tag{6}$$

$$\oint_{\mathcal{S}} v^{\mathbf{i}}\mu\,\mathrm{d}\mathcal{S}_1 = o\big(r^{-4}\big)\,. \tag{7}$$

Here v^a is the outward directed unit normal to the large sphere $\mathcal{S}$ of radius r in the hyperplane Σ, and $\mathrm{d}\mathcal{S}_1$ is the area element on the *unit* sphere. However, the *global integral conditions* (6)–(7) are only implicit restrictions on the asymptotic behaviour of μ and j^a, and hence it is difficult to use them in practice. If we are not interested in the exact boundary conditions, as in the present discussion, then we prefer to have only an *explicitly given* sufficient condition. Such a sufficient condition might be the global *parity condition*: The leading terms in (4) and (5) are required to be even parity functions of their second argument: $\mu^{(4)}(t, \frac{x^{\mathbf{k}}}{r}) = \mu^{(4)}(t, -\frac{x^{\mathbf{k}}}{r})$ and $j^{\mathbf{i}(4)}(t, \frac{x^{\mathbf{k}}}{r}) = j^{\mathbf{i}(4)}(t, -\frac{x^{\mathbf{k}}}{r})$. Then the fall-off and parity conditions together ensure the finiteness of the total energy-momentum and (relativistic) angular momentum of the matter fields.

It is easy to check that if the fall-off and parity conditions above are imposed not only on a single spacelike hyperplane but on boosted hyperplanes as well, then the spatial stress part of the energy-momentum tensor, $\sigma^{ab} := P^a_c P^b_d T^{cd}$, must also have the asymptotic structure

$$\sigma^{\mathbf{ij}} = \frac{1}{r^4}\sigma^{\mathbf{ij}(4)}\Big(t, \frac{x^{\mathbf{k}}}{r}\Big) + o\big(r^{-4}\big)\,, \tag{8}$$

and the leading term $\sigma^{\mathbf{ij}(4)}(t, \frac{x^{\mathbf{k}}}{r})$ must be an even parity function of $\frac{x^{\mathbf{k}}}{r}$.

2.4 Asymptotically Cartesian Coordinate Systems

By the results of the previous two subsections the Cartesian coordinates appear to play a fundamental role in the definition and the study of the properties of the conserved quantities in Minkowski spacetime. But as we saw in Subsect. 2.1, the Cartesian coordinates have metrical content, because, by their very definition, they are adapted to exact geometric symmetries of the spacetime. However, primarily we are interested in general, non-flat asymptotically flat spacetimes, where we do not have any exact geometric symmetry. Thus the question arises naturally whether or not there is some natural generalization of the familiar Cartesian coordinates, at least asymptotically, even in a general asymptotically flat spacetime, which could play an analogous role in constructing the conserved quantities.

Such an asymptotically Cartesian coordinate system $(\tau, \eta^{\mathbf{i}})$ may be based on a foliation Σ_τ of the asymptotically flat spacetime, which foliation can be characterized on a typical leaf Σ of the foliation by the lapse function N. Furthermore, we need to have a shift vector N^a as well, which tells us how the spatial Cartesian coordinates $\eta^{\mathbf{i}}$, introduced on one leaf of the foliation, is extended to the neighbouring leaves. Thus we would like to find a criterion, formulated in terms of the lapse and the shift, when to consider the corresponding coordinate system $(\tau, \eta^{\mathbf{i}})$ to be asymptotically Cartesian.

In Minkowski spacetime the lapse of the Cartesian coordinate system is the constant function with value 1, and the shift is identically vanishing. Therefore, it seems natural to consider the coordinate system $(\tau, \eta^{\mathbf{i}})$ to be asymptotically Cartesian only if $N \to 1$ and $N^a \to 0$ at infinity uniformly, independently of the direction in which the limit is taken on Σ. This naive criterion can also be supported by a formal analysis of the coordinate systems in the conformally compactified Minkowski spacetime near the spatial infinity i^0 [14]: There exists a flat metric ${}_0q_{ab}$ on Σ such that $q_{\mathbf{ij}} - {}_0q_{\mathbf{ij}}$ and $\chi_{\mathbf{ij}}$, the components of the difference of the induced and the flat metrics and of the extrinsic curvature in the ${}_0q_{ab}$–Cartesian coordinates $\eta^{\mathbf{i}}$, respectively, tend to zero as $R^2 := \delta_{\mathbf{ij}}\eta^{\mathbf{i}}\eta^{\mathbf{j}}$ tends to infinity, and moreover $N(\tau, \eta^{\mathbf{k}}) = 1 + O(R^{-1})$ and $N^{\mathbf{i}}(\tau, \eta^{\mathbf{k}}) = O(R^{-1})$. In fact, an asymptotically vanishing N would correspond to a foliation in which the time separation of the different leaves tends to zero, while an asymptotically diverging N would correspond to one in which this time separation is diverging. Thus, in particular, in Minkowski spacetime the coordinate transformation connecting the Cartesian coordinate system to a system $(\tau, \eta^{\mathbf{i}})$ based on an asymptotically vanishing lapse is getting to be singular, i.e. $(\tau, \eta^{\mathbf{i}})$ is "collapsing" asymptotically.

If $(\tau, \eta^{\mathbf{i}})$ is an asymptotically Cartesian coordinate system in Minkowski spacetime based on a smooth spacelike Cauchy surface Σ extending to the spatial infinity, then the Killing field (1) takes the form

$$K_e = M_{\mathbf{ij}}\big(\eta^{\mathbf{i}} D_e\eta^{\mathbf{j}} - \eta^{\mathbf{j}} D_e\eta^{\mathbf{i}}\big) + 2M_{\mathbf{i0}}\big(\eta^{\mathbf{i}}\tau_e - \tau D_e\eta^{\mathbf{i}}\big) + s_{\mathbf{i}} D_e\eta^{\mathbf{i}} + s\tau_e \,. \qquad (9)$$

Here D_e is the intrinsic derivative operator, τ_e the future pointing unit timelike normal to Σ and $s(\tau,\eta^{\mathbf{k}}) = s^{(0)}(\tau,\frac{\eta^{\mathbf{k}}}{R}) + O(R^{-1})$, $s_{\mathbf{i}}(\tau,\eta^{\mathbf{k}}) = s^{(0)}_{\mathbf{i}}(\tau,\frac{\eta^{\mathbf{k}}}{R}) + O(R^{-1})$. Thus s and $s_e := s_{\mathbf{i}} D_e \eta^{\mathbf{i}}$, which would be the time and space translation parts of K_e in (1), respectively, depend on τ, $\eta^{\mathbf{i}}$ and $1/R$. Therefore, they are analogous to the supertranslations of the cuts of future null infinity, and the proper translations correspond only to special supertranslations. We will see in Subsects. 3.2 and 4.1 that these are precisely the $\eta^{\mathbf{i}}$-independent supertranslations, while those that are odd parity functions of $\frac{\eta^{\mathbf{k}}}{R}$ are the proper supertranslations and have only gauge content.

2.5 Conservation Properties

We saw in Subsect. 2.2 that the quasi-local energy-momentum and angular momentum are conserved with respect to a time evolution characterized by a vector field ξ^a if the evolution preserves $D(\Sigma)$, i.e. the lapse part of ξ^a is vanishing on $\mathcal{S}$ and the shift part is tangent to $\mathcal{S}$ on $\mathcal{S}$. In the present subsection we formulate the analogous question for the total energy-momentum and angular momentum.

Thus let Σ_τ be a foliation of the Minkowski spacetime by smooth Cauchy surfaces, let t^a be its future pointing unit timelike normal and N the lapse of the foliation. Let N^a be the shift vector and define the "general time axis" $\xi^a := Nt^a + N^a$. Then we can take the integrals (3) defining the total energy-momentum and angular momentum on the leaves Σ_τ and calculate their Lie derivative along ξ^a. Our question is what asymptotic conditions should the lapse and the shift satisfy such that these Lie derivatives be vanishing. However, this analysis consists of two things. The first is that even though the integral (3) on a specific hypersurface is finite, it is not necessarily finite on the hypersurfaces obtained by "time evolution" along ξ^a; i.e. *we should ensure that the boundary conditions ensuring the finiteness of (3) be preserved*. The second is to ensure that these finite integrals be the same.

Nevertheless, this analysis can be, and in the next section will be, carried out even in general asymptotically flat spacetimes with vector fields K^a having the asymptotic structure more general than (9). We will see that the total energy-momentum and (relativistic) angular momentum are conserved even if N and N^a are linearly diverging (see (21)–(22)).

3 Asymptotically Flat Spacetimes

3.1 The Boundary Conditions

The definition of the asymptotic flatness of a spacetime that we adopt in the present paper is probably the oldest one. We say that a spacetime is asymptotically flat at spatial infinity if it contains an asymptotically flat spacelike

hypersurface Σ. Thus we should define the asymptotic flatness of such a hypersurface. We say that the spacelike hypersurface Σ is (k,l)-asymptotically flat, if (1) there is a (negative definite) background metric ${}_0q_{ab}$ on Σ, which is flat outside a large compact subset $K \subset \Sigma$ such that $\Sigma - K$ is diffeomorphic to $\mathbb{R}^3$ minus a solid ball; (2) for some positive k and l the components $q_{\mathbf{ij}}$ and $\chi_{\mathbf{ij}}$ of the physical induced metric and of the extrinsic curvature, respectively, in the ${}_0q_{ab}$-Cartesian coordinate system on $\Sigma - K$ satisfy the fall-off conditions

$$q_{\mathbf{ij}} - {}_0q_{\mathbf{ij}} = \frac{1}{r^k} q^{(k)}_{\mathbf{ij}} + o\big(r^{-k}\big), \tag{10}$$

$$\chi_{\mathbf{ij}} = \frac{1}{r^l} \chi^{(l)}_{\mathbf{ij}} + o\big(r^{-l}\big)\ ; \tag{11}$$

and, (3) the leading terms $q^{(k)}_{\mathbf{ij}}$ and $\chi^{(l)}_{\mathbf{ij}}$ are even and odd parity functions of $\frac{x^{\mathbf{k}}}{r}$, respectively. Here r is the radial coordinate defined by $r^2 := \delta_{\mathbf{ij}} x^{\mathbf{i}} x^{\mathbf{j}}$. In general these conditions do not imply that every component e.g. of the derivative ${}_0D_c q_{ab}$ tends to zero as $1/r^{k+1}$, where ${}_0D_c$ is the derivative operator determined by the background metric, which would be a useful property in practice. Similarly, still not every component of ${}_0D_c\chi_{ab}$ tends to zero as $1/r^{l+1}$. If, however, we assume that the "rests" $m_{ab} := q_{ab} - {}_0q_{ab} - r^{-k} q^{(k)}_{ab}$ and $k_{ab} := \chi_{ab} - r^{-l}\chi^{(l)}_{ab}$ also satisfy

$$ {}_0D_c m_{ab} = o\big(r^{-k-1}\big),\ \ {}_0D_{d0}D_c m_{ab} = o\big(r^{-k-2}\big),\ \ \ldots \tag{12}$$

$$ {}_0D_c \chi_{ab} = o\big(r^{-l-1}\big),\ \ {}_0D_{d0}D_c \chi_{ab} = o\big(r^{-l-2}\big),\ \ \ldots \tag{13}$$

then, together with (10) and (11), these imply ${}_0D_{e_1}\ldots{}_0D_{e_s} q_{ab} = O(r^{-k-s})$ and ${}_0D_{e_1}\ldots{}_0D_{e_s}\chi_{ab} = O(r^{-l-s})$ for any $s = 1, 2, \ldots$, and the parity of these derivatives is $(-)^s$ and $(-)^{s+1}$, respectively. The properties $m_{ab} = o(r^{-k})$, ${}_0D_e m_{ab} = o(r^{-k-1})$, $\ldots$ of m_{ab} will be denoted by $m_{ab} = o^\infty(r^{-k})$. Although it would be enough to require ${}_0D_{e_1}\ldots{}_0D_{e_s} m_{ab} = o(r^{-k-s})$ only for some finite s depending on the order of the derivatives that appears in the actual calculations, for the sake of simplicity we assume that $m_{ab} = o^\infty(r^{-k})$. Similarly, we require that $k_{ab} = o^\infty(r^{-l})$.

We assume that the matter fields satisfy boundary conditions that yield energy density μ, momentum density j^a and spatial stress σ^{ab} satisfying the fall-off and parity conditions that we discussed in Subsect. 2.3, defined with respect to the ${}_0q_{ab}$-Cartesian coordinate system. Furthermore, again by technical reasons, we assume that the "rests" appearing in (4), (5) and (8) are also $o^\infty(r^{-4})$. Then we can form the integral

$$\mathtt{Q}^m\big[M, M^a\big] := \int_\Sigma \big(Mt_a + M_a\big) T^{ab} t_b \mathrm{d}\Sigma\ , \tag{14}$$

and, as a consequence of the boundary conditions for μ and j^a, this integral exists if the asymptotic form of M and M^a is given by

$$M\big(t,x^{\mathbf{k}}\big)=r^A M^{(A)}\Big(t,\frac{x^{\mathbf{k}}}{r}\Big)+o^\infty\big(r^A\big), \tag{15}$$

$$M_{\mathbf{i}}\big(t,x^{\mathbf{k}}\big)=r^B M^{(B)}_{\mathbf{i}}\Big(t,\frac{x^{\mathbf{k}}}{r}\Big)+o^\infty\big(r^B\big)\ , \tag{16}$$

where $A,B\leq 1$ and if the equality holds in these inequalities then $M(t,\frac{x^{\mathbf{k}}}{r})$ and $M_{\mathbf{i}}(t,\frac{x^{\mathbf{k}}}{r})$, respectively, must be odd parity functions of $\frac{x^{\mathbf{k}}}{r}$. Note that (9), and hence (1) also, are special cases of (15)–(16). In the next subsection we discuss the time dependence of $\mathtt{Q}^m[M,M^a]$.

3.2 The Evolution Equations

Let the spacetime be foliated by smooth spacelike Cauchy hypersurfaces Σ_t, and let a "general time axis" $\xi^a=Nt^a+N^a$ be also given. Then the $3+1$ form of the equation $T^{ab}{}_{;b}=0$ is well known to be

$$\dot\mu=N\Big(-D_aj^a+\sigma^{ab}\chi_{ab}-\frac{2}{N}j^aD_aN-\mu\chi\Big)+\text{Ł}_{\mathbf{N}}\mu, \tag{17}$$

$$\dot j^b=N\Big(-D_a\sigma^{ab}-\frac{1}{N}\sigma^{ba}D_aN+\mu\frac{1}{N}D^bN-2j_a\chi^{ba}-\chi j^b\Big)+\text{Ł}_{\mathbf{N}}j^b\ , \tag{18}$$

where the dot denotes the projection to the leaves of the foliation of the Lie derivative along ξ^a. They describe the evolution of the energy density and the momentum density of the matter fields along the integral curves of ξ^a. Similarly, the evolution equations for the geometry are

$$\dot q_{ab}=2N\chi_{ab}+\text{Ł}_{\mathbf{N}}q_{ab}, \tag{19}$$

$$\begin{aligned}\dot\chi_{ab}=&N\Big(-R_{ab}+2\chi_{ac}\chi^c{}_b-\chi\chi_{ab}\Big)+\text{Ł}_{\mathbf{N}}\chi_{ab}-D_aD_bN\\&+\Lambda Nq_{ab}+\kappa N\Big(-\sigma_{ab}+\frac{1}{2}\sigma^e{}_eq_{ab}+\frac{1}{2}\mu q_{ab}\Big)\ .\end{aligned} \tag{20}$$

The first is a simple consequence of the definitions, but the second is the space-space projection of the Einstein equations.

Next suppose that the spacetime is asymptotically flat (whenever the cosmological constant Λ is zero), and characterize the foliation and the general time axis on a typical Cauchy surface Σ by the lapse N and the shift N^a. In the previous subsection we defined the asymptotic flatness of the spacetime by the existence of an appropriately defined asymptotically flat spacelike hypersurface. However, the existence of such a single hypersurface does not imply that the evolution of such a hypersurface will be asymptotically flat, i.e. the boundary conditions are not necessarily preserved by the dynamical equations. Thus our question is what conditions should we impose on the lapse and the shift such that the evolution equations (17)–(20) preserve the fall-off and parity conditions, both for the matter fields and the geometry.

Assuming that the lapse and the shift have the a priori asymptotic form $N(t,x^{\mathbf{k}})=r^CN^{(C)}(t,\frac{x^{\mathbf{k}}}{r})+o^\infty(r^C)$ and $N_{\mathbf{i}}(t,x^{\mathbf{k}})=r^DN^{(D)}_{\mathbf{i}}(t,\frac{x^{\mathbf{k}}}{r})+o^\infty(r^D)$

for some C and D, we can evaluate the right hand side of the evolution equations. If we require that the leading orders and parities on both sides coincide, we obtain two results. The first is a link between the fall-off rates for the metric and the extrinsic curvature: In the generic case $l = k+1$. (For the exceptional cases see [14].) The other is the detailed asymptotic structure of the lapse and the shift, given by

$$N\big(t, x^{\mathbf{k}}\big) = 2x^{\mathbf{k}}\beta_{\mathbf{k}}\big(t\big) + \tau\big(t\big) + r^E\nu^{(E)}\Big(t, \frac{x^{\mathbf{k}}}{r}\Big) + o^\infty\big(r^E\big), \tag{21}$$

$$N_{\mathbf{i}}\big(t, x^{\mathbf{k}}\big) = 2x^{\mathbf{k}}\rho_{\mathbf{ki}}\big(t\big) + \tau_{\mathbf{i}}\big(t\big) + r^F\nu^{(F)}_{\mathbf{i}}\Big(t, \frac{x^{\mathbf{k}}}{r}\Big) + o^\infty\big(r^F\big)\ . \tag{22}$$

Here the coefficients $\beta_{\mathbf{k}}(t)$, $\tau(t)$, $\rho_{\mathbf{ki}}(t)$ and $\tau_{\mathbf{i}}(t)$ are arbitrary functions of t, the powers E and F are bounded from above by the fall-off rate of the metric: $E, F \leq (1-k)$, and if the equality holds in these inequalities then the functions $\nu^{(E)}(t, \frac{x^{\mathbf{k}}}{r})$ and $\nu^{(F)}_{\mathbf{i}}(t, \frac{x^{\mathbf{k}}}{r})$ are odd parity functions of their second argument, respectively.

Since the structure of $N_{\mathbf{i}}$ is similar to that of N, it is enough to discuss only e.g. (21). By $k > 0$ the leading term in (21) is the first, but to decide whether the next order is the second or the third, we should consider the disjoint cases $k > 1$, $k < 1$ and $k = 1$. If $k > 1$, which corresponds to a fast fall-off metric, then the third term tends to zero at infinity as r^E, where E is *negative*, whenever the next order term is the second. If $k < 1$, which corresponds to a slow fall-off, E may be positive, and if E is actually positive, then the third term is diverging. In this case there is no reason to keep the second term, because that cannot be isolated in the presence of the uncontrollable diverging term. If $k = 1$, then E may be zero, and if it is actually zero, then both the second and the third terms are asymptotically of the same order. However, in spite of the fact that the third term is uncontrollable and of the same order asymptotically as the second, we can make a natural distinction between these: The second, being independent of the spatial coordinates, is an *even* parity, while the third is an *odd* parity function of $\frac{x^{\mathbf{k}}}{r}$. Thus the structure of N and N^a resembles the structure of the timelike and spacelike projection of the Killing fields of the Minkowski spacetime given by (1) or rather (9). In particular, $\beta_{\mathbf{k}}(t)$, $\rho_{\mathbf{ki}}(t)$, $\tau(t)$ and $\tau_{\mathbf{i}}(t)$ are analogous to the boost, rotation, time translation and spatial translation generators, and the terms $r^E\nu^{(E)}$ and $r^F\nu^{(F)}_{\mathbf{i}}$ are similar to the proper temporal and spatial supertranslations of (9). However, while the components of the Killing fields have a special time dependence, the parameters $\beta_{\mathbf{k}}(t)$, $\rho_{\mathbf{ki}}(t)$, $\tau(t)$ and $\tau_{\mathbf{i}}(t)$ may have arbitrary time dependence.

Defining the integral $\mathbb{Q}^m[M, M^a]$, given by (14), on each of the leaves Σ_t of the foliation, one can compute its time derivative. It is

$$\dot{\mathtt{Q}}^m[M,M^a] = \int_{\Sigma_t}\Big(\mu\big(\dot{M}+M^aD_aN-N^aD_aM\big) + j_a\big(\dot{M}^a+ND^aM-MD^aN-[N,M]^a\big) + \sigma_{ab}N\big(M\chi^{ab}+D^{(a}M^{b)}\big) + D_a\big(\big(\mu M+j_bM^b\big)N^a-\big(j^aM+\sigma^{ab}M_b\big)N\big)\Big)\mathrm{d}\Sigma_t\,. \quad (23)$$

Taking into account the boundary conditions and substituting the asymptotic form (21)–(22) here we find that $\dot{\mathtt{Q}}^m[M,M^a]$ is finite (such that the integral of the total divergence in (23) is zero). We will see in Subsect. 5.1 that the coefficients of μ, j_a and σ_{ab} in the volume integral of (23) are precisely the various 3+1 parts of the Killing operator $\nabla^{(a}K^{b)}$ acting on $K^a := Mt^a+M^a$. Thus for Killing vectors $\mathtt{Q}^m[M,M^a]$ is constant in time even if the "time evolution" is defined by $\xi^a = Nt^a+N^a$ with asymptotically linearly diverging N and N^a given by (21)–(22).

The question of whether the evolution equations preserve the boundary conditions was investigated first by Beig and Ó Murchadha [5]. However, they considered only the vacuum equations with the $1/r$ and $1/r^2$ a priori fall-off of the metric and extrinsic curvature, respectively, and they assumed a priori that the lapse and the shift are time independent. While the first two are not serious limitations of their investigations, we do not see any reason to assume the time independence of N and N^a. In fact, the evolution equations allow their arbitrary time dependence, and, as we will see, the assumption of their time independence is too restrictive and we should abandon this.

Finally, for later convenience, it seems natural to introduce two notations here. We will denote by $\mathcal{A}$ the set of all the pairs (N,N^a) of lapses and shifts with the asymptotic form (21)–(22). Such pairs may be called the "allowed time axes", and obviously $\mathcal{A}$ can be endowed with a natural real vector space structure. We denote by $\mathcal{G}$ the subspace of $\mathcal{A}$ consisting of those pairs in which the 'parameters' $\beta_{\mathbf{k}}(t)$, $\rho_{\mathbf{ki}}(t)$, $\tau(t)$ and $\tau_{\mathbf{i}}(t)$ are all vanishing identically. We will see in the next subsection that, for $k \geq 1$, the generators of the gauge transformations in the phase space of vacuum general relativity are precisely the elements of $\mathcal{G}$. Thus we refer to $\mathcal{G}$ as to the space of the gauge generators even for $k > 0$.

4 The Hamiltonian Phase Space of Vacuum GR

4.1 The Phase Space and the General Beig–Ó Murchadha Hamiltonian

The configuration space $\mathcal{Q}$ for the asymptotically flat spacetimes is the set of the (negative definite) metrics on the 3-manifold Σ, a typical spacelike Cauchy surface in spacetime, satisfying the fall-off and parity conditions of Subsect. 3.1. Recalling that a curve in $\mathcal{Q}$ is a smooth 1-parameter family of

metrics $q_{ab}(u)$ and the tangent vector of this curve at the point $q_{ab} := q_{ab}(0) \in \mathcal{Q}$ is defined to be the derivative $\delta q_{ab} := (\mathrm{d}q_{ab}(u)/\mathrm{d}u)|_{u=0}$, the tangent vector δq_{ab} satisfies the same boundary conditions as the metric q_{ab} itself does. The space of the tangent vectors at q_{ab} is denoted by $T_{q_{ab}}\mathcal{Q}$. Recall also that a 1-form at the point $q_{ab} \in \mathcal{Q}$ is a symmetric tensor density on Σ, which, at the same time, is a linear mapping $\tilde{p}^{ab} : T_{q_{ab}}\mathcal{Q} \to \mathbb{R}$ defined explicitly by $\langle \tilde{p}^{ab}, \delta q_{ab}\rangle := \int_\Sigma \tilde{p}^{ab}\delta q_{ab}\mathrm{d}^3x$. However, the requirement that its action on the tangent vectors be finite restricts its asymptotic structure. Indeed, if we write $\tilde{p}^{ab} = \frac{1}{r^m}\tilde{p}^{(m)ab} + o(r^{-m})$ for some $m > 0$, then from $\langle \tilde{p}^{ab}, \delta q_{ab}\rangle < \infty$ we obtain that $m \geq 3 - k$, and if the equality holds in this inequality then the components $\tilde{p}^{(m)\mathbf{ij}}$ of the leading term must be odd parity functions of $\frac{x^{\mathbf{k}}}{r}$. The space of these 1-forms at q_{ab}, the cotangent space of $\mathcal{Q}$ at q_{ab}, is denoted by $T^*_{q_{ab}}\mathcal{Q}$.

The phase space of vacuum general relativity is the cotangent bundle $T^*\mathcal{Q} := \{(\tilde{p}^{ab}, q_{ab})| +\ \text{boundary conditions}\}$ of the configuration space with its natural symplectic structure: If $\mathcal{X} := (\delta\tilde{p}^{ab}, \delta q_{ab})$ and $\mathcal{X}' := (\delta'\tilde{p}^{ab}, \delta' q_{ab})$ are any two tangent vectors at some point $(\tilde{p}^{ab}, q_{ab}) \in T^*\mathcal{Q}$, then let $2\Omega_{(\tilde{p}^{ab}, q_{ab})}(\mathcal{X}, \mathcal{X}') := \int_\Sigma(\delta\tilde{p}^{ab}\delta' q_{ab} - \delta'\tilde{p}^{ab}\delta q_{ab})\mathrm{d}^3x$. Then the boundary conditions for the metrics and the canonical momenta ensure that $\Omega(\mathcal{X}, \mathcal{X}')$ is already finite.

On the other hand, the canonical momentum $\tilde{p}^{ab}$ is well known to be the expression

$$\tilde{p}^{ab} = \frac{1}{2\kappa}\sqrt{|q|}\Big(\chi^{ab} - \chi q^{ab}\Big) = \frac{1}{r^{k+1}}\tilde{P}^{(k+1)ab} + o^\infty\big(r^{-k-1}\big) \tag{24}$$

of the metric and the extrinsic curvature, where we gave its asymptotic expansion too. Here the components of $\tilde{P}^{(k+1)ab}$ in the ${}_0q_{ab}$-Cartesian coordinates are odd parity functions of $\frac{x^{\mathbf{k}}}{r}$. Therefore, comparing this fall-off rate with the condition $m \geq 3-k$ obtained above, we find that *the applicability of the basic concepts of the symplectic framework already excludes the slow fall-off metrics*, i.e. $k \geq 1$ must be assumed. Thus the a priori fall-off $1/r$ considered by Beig and Ó Murchadha is the slowest possible in the symplectic framework.

Four of the vacuum Einstein equations, ${}^4G_{ab}t^at^b = 0$ and ${}^4G_{bc}P^b_at^c = 0$, play the role of constraints in the initial value as well as in the Hamiltonian formulation of the theory. In the phase space context they are represented by the vanishing of the so-called constraint functions

$$C\big[\nu, \nu^a\big] := \int_\Sigma \left(-\frac{1}{2\kappa}\left(R + \frac{4\kappa^2}{|q|}\left[\frac{1}{2}\tilde{p}^2 - \tilde{p}^{ab}\tilde{p}_{ab}\right]\right)\sqrt{|q|}\nu - 2\big(D_a\tilde{p}^{ab}\big)\nu_b\right)\mathrm{d}^3x\ , \tag{25}$$

parameterized by pairs (ν, ν^a) of functions and vector fields, which may be functions of the external time coordinate as well. A tedious but straightforward calculation shows that *the constraint functions are finite and functionally differentiable with respect to the canonical variables on the whole phase space and close to a Lie algebra if and only if* $(\nu, \nu^a) \in \mathcal{G}$. Since, via

the symplectic 2-form, they generate gauge motions in the constraint surface $\Gamma \subset T^*\mathcal{Q}$, $\mathcal{G}$ *can be identified with the space of the infinitesimal gauge generators of Einstein's theory of the vacuum asymptotically flat spacetimes.*

The dynamics in the phase space is generated by the Hamiltonian, whose general form is the sum of a constraint function and the integral of an appropriately chosen total divergence. This total divergence should be chosen in such a way that the corresponding Hamilton equations be just the correct evolution equations (19) and (20) [4]. Beig and Ó Murchadha [5] showed that *the Hamiltonian*

$$\begin{aligned} H\big[M,M^a\big] := C\big[M,M^a\big] + \int_\Sigma 2D_a\big(\tilde{p}^{ab}M_b\big)\mathrm{d}^3x \\ -\frac{1}{2\kappa}\int_\Sigma D_a\Big(\, Mq^{ab}q^{cd}\big({}_0D_cq_{bd} - {}_0D_bq_{cd}\big) \\ + \big({}_0D_bM\big)q^{ab}q^{cd}\big(q_{cd} - {}_0q_{cd}\big) \\ - \big({}_0D_cM\big)q^{ab}q^{cd}\big(q_{bd} - {}_0q_{bd}\big)\Big)\sqrt{|q|}\mathrm{d}^3x \end{aligned} \tag{26}$$

is finite and functional differentiable with respect to the canonical variables on the whole phase space and close to a Lie algebra if and only if $(M,M^a)\in\mathcal{A}$. Thus we call H given by (26) the Beig–Ó Murchadha Hamiltonian. Note that M and M^a need not be time independent, they may still have arbitrary time dependence. The Poisson bracket of two Beig–Ó Murchadha Hamiltonians, parameterized by (M,M^a) and $(\bar{M},\bar{M}^a)$, respectively, is

$$\begin{aligned} &\Big\{H\big[M,M^a\big],H\big[\bar{M},\bar{M}^a\big]\Big\} \\ &\quad = -H\Big[\text{Ł}_{\mathbf{M}}\bar{M} - \text{Ł}_{\bar{\mathbf{M}}}M\ ,\ \big[M,\bar{M}\big]^a - \big(MD^a\bar{M} - \bar{M}D^aM\big)\Big]\,. \end{aligned} \tag{27}$$

Furthermore, for infinitesimal gauge generators the Hamiltonian of Beig and Ó Murchadha reduces to a constraint function: $H[\nu,\nu^a] = C[\nu,\nu^a]$. Therefore, *the Beig–Ó Murchadha Hamiltonians, parameterized by the elements of* $\mathcal{A}$*, form a Poisson algebra* $\mathcal{H}$*, in which the constraints, parameterized by the elements of* $\mathcal{G}$*, form an ideal* $\mathcal{C}$. The quotient $\mathcal{H}/\mathcal{C}$, which is again a Lie algebra, is the set of the Hamiltonians modulo the "gauge transformations". However, this quotient Lie algebra is spanned by the time dependent parameters $\beta_{\mathbf{k}}(t)$, $\rho_{\mathbf{ki}}(t)$, $\tau(t)$ and $\tau_{\mathbf{i}}(t)$, and hence it is infinite dimensional.

4.2 Physical Quantities from the Beig–Ó Murchadha Hamiltonians with Time-Independent Lapses and Shifts

As we mentioned, Beig and Ó Murchadha concentrated on the $k = 1$ case and assumed that M and M^a were time independent:

$$M\big(x^{\mathbf{k}}\big) = 2x^{\mathbf{k}}B_{\mathbf{k}} + T + \nu^{(0)}\Big(\frac{x^{\mathbf{k}}}{r}\Big) + o^\infty\big(r^{-0}\big)\,, \tag{28}$$

$$M_{\mathbf{i}}\big(x^{\mathbf{k}}\big) = 2x^{\mathbf{k}}R_{\mathbf{ki}} + T_{\mathbf{i}} + \nu^{(0)}_{\mathbf{i}}\Big(\frac{x^{\mathbf{k}}}{r}\Big) + o^\infty\big(r^{-0}\big)\,. \tag{29}$$

Here $B_{\mathbf{k}}$, $R_{\mathbf{ki}}$, T and $T_{\mathbf{i}}$ are real constants. The space of such pairs (M, M^a) will be denoted by ${}_0\mathcal{A}$, and the subspace of the "infinitesimal time independent gauge generators" by ${}_0\mathcal{G}$. Then the Beig–Ó Murchadha Hamiltonians parameterized by the elements of ${}_0\mathcal{A}$ form a Poisson algebra ${}_0\mathcal{H}$, in which the constraints parameterized by the elements of ${}_0\mathcal{G}$ form a Lie ideal ${}_0\mathcal{C}$. The quotient Lie algebra ${}_0\mathcal{H}/{}_0\mathcal{C}$ is spanned by the ten real parameters $B_{\mathbf{k}}$, $R_{\mathbf{ki}}$, T and $T_{\mathbf{i}}$, and, as Beig and Ó Murchadha showed, this is isomorphic to the Poincaré Lie algebra.

Since $H[M, M^a]$ is linear in M and M^a, its restriction to the constraint surface Γ is a 2-surface integral at infinity of the boundary expression in (26), which involves the parameters T, $T_{\mathbf{i}}$, $R_{\mathbf{ki}}$ and $B_{\mathbf{k}}$ linearly. The coefficients of these parameters define the total energy, linear momentum, spatial angular momentum and centre-of-mass, respectively:

$$H\big[M, M^a\big]|_\Gamma =: T\mathtt{P}^{\mathbf{0}} + T_{\mathbf{i}}\mathtt{P}^{\mathbf{i}} + R_{\mathbf{ij}}\mathtt{J}^{\mathbf{ij}} + 2B_{\mathbf{i}}\mathtt{J}^{\mathbf{i0}} \ . \tag{30}$$

The total energy and linear momentum defined in this way is precisely the familiar ADM energy and linear momentum [3], and the spatial angular momentum is just the angular momentum of Regge and Teitelboim [4]. However, the centre-of-mass expression deviates slightly from that given by Regge and Teitelboim. While the Regge–Teitelboim centre-of-mass is not always finite, the expression given by the Beig–Ó Murchadha Hamiltonian is. We call the latter expression the Beig–Ó Murchadha centre-of-mass.

4.3 Transformation and Conservation Properties

We saw in Subsect. 2.2 that even the quasi-locally defined energy-momentum and (relativistic) angular momentum of the matter fields transform in the correct way under the Poincaré transformations of the Cartesian coordinates. Since these transformations can also be interpreted as the action of the symmetries of the Minkowski spacetime, it is natural to ask about the transformation properties of the total energy, linear momentum, spatial angular momentum and centre-of-mass, introduced in the previous subsection, under the action of the "asymptotic symmetries" of the spacetime. Roughly, the structure of (28) and (29) is similar to the structure of the time and space projections of the Killing fields (1), thus it seems natural to identify them as the "asymptotic symmetry generators". Hence we would have to define the action of them on the physical quantities in question.

However, since $\mathtt{P}^{\mathbf{0}}$, $\mathtt{P}^{\mathbf{i}}$, $\mathtt{J}^{\mathbf{ij}}$ and $\mathtt{J}^{\mathbf{i0}}$ were introduced in the phase space rather than the spacetime, one may think that it is enough to clarify their transformation properties in the *phase space*. To do this we need an implementation of the "asymptotic symmetry generators" in the phase space in the form of some functionally differentiable function. However, we already do have such an implementation, namely the Beig–Ó Murchadha Hamiltonian parameterized by the 'symmetry generators', and hence we can define its action. The action of the "symmetry generator" $(\bar{M}, \bar{M}^a) \in {}_0\mathcal{A}$ on the total

energy, linear momentum, spatial angular momentum and centre-of-mass is defined by the value on the constraint surface Γ of the Poisson bracket of the Hamiltonian implementing the 'asymptotic symmetry' and the Hamiltonian defining the physical quantities. Formally it is

$$\delta_{(\bar{M},\bar{M}^a)}\Big(T\mathtt{P}^{\mathbf{0}}+T_{\mathbf{i}}\mathtt{P}^{\mathbf{i}}+R_{\mathbf{ij}}\mathtt{J}^{\mathbf{ij}}+2B_{\mathbf{i}}\mathtt{J}^{\mathbf{i0}}\Big) := \Big\{H\big[\bar{M},\bar{M}^a\big],H\big[M,M^a\big]\Big\}|_{\Gamma}\ . \quad (31)$$

Evaluating the right hand side of (31) by using (27), the result can be summarized as follows: If we form the column vectors

$$\mathtt{P}^{\underline{a}} := \begin{pmatrix} \mathtt{P}^{\mathbf{0}} \\ \mathtt{P}^{\mathbf{i}} \end{pmatrix}\ , \qquad \bar{c}^{\underline{a}} := \begin{pmatrix} \bar{T}^{\mathbf{0}} \\ \bar{T}^{\mathbf{i}} \end{pmatrix}\ ,$$

and the 4×4 anti-symmetric matrices

$$\mathtt{J}^{\underline{a}\,\underline{b}} := \begin{pmatrix} 0 & -\mathtt{J}^{\mathbf{i0}} \\ \mathtt{J}^{\mathbf{i0}} & \mathtt{J}^{\mathbf{ij}} \end{pmatrix}\ , \qquad \bar{\lambda}_{\underline{a}\,\underline{b}} := \begin{pmatrix} 0 & -2\bar{B}_{\mathbf{j}} \\ 2\bar{B}_{\mathbf{i}} & 2\bar{R}_{\mathbf{ij}} \end{pmatrix}\ ,$$

then we obtain

$$\delta_{(\bar{M},\bar{M}^e)}\mathtt{P}^{\underline{a}} = -\mathtt{P}^{\underline{b}}\,\bar{\lambda}_{\underline{b}}{}^{\underline{a}} \quad (32)$$

$$\delta_{(\bar{M},\bar{M}^e)}\mathtt{J}^{\underline{a}\,\underline{b}} = -\Big(\mathtt{J}^{\underline{c}\,\underline{b}}\,\bar{\lambda}_{\underline{c}}{}^{\underline{a}} + \mathtt{J}^{\underline{a}\,\underline{c}}\,\bar{\lambda}_{\underline{c}}{}^{\underline{b}} + \big(\bar{c}^{\underline{a}}\,\mathtt{P}^{\underline{b}} - \bar{c}^{\underline{b}}\,\mathtt{P}^{\underline{a}}\big)\Big)\ . \quad (33)$$

This is precisely (minus) the action of the infinitesimal Poincaré transformation, parameterized by $\bar{c}^{\underline{a}} \in \mathbb{R}^4$ and the Lorentz Lie algebra element $\bar{\lambda}_{\underline{b}}{}^{\underline{a}}$, on an energy-momentum 4-vector and a relativistic angular momentum 4-tensor. Therefore, the total energy, linear momentum, spatial angular momentum and the Beig–Ó Murchadha centre-of-mass form Lorentz-covariant quantities, and transform *in the phase space* in the correct way.

The next issue that we should discuss is whether these quantities are conserved in time, or, more generally, under what conditions on the lapse and shift defining the time evolution do we have conserved total energy-momentum and (relativistic) angular momentum. Thus let $(N,N^a)\in\mathcal{A}$ be any allowed (maybe time dependent) time axis, given explicitly by (21) and (22) with $k=1$. Then we define the time derivative of $\mathtt{P}^{\underline{a}}$ and $\mathtt{J}^{\underline{a}\,\underline{b}}$ by the value on the constraint surface of the Poisson bracket of the Hamiltonian defining the time evolution via the dynamical equations and the Hamiltonian defining the physical quantities:

$$\frac{\mathrm{d}}{\mathrm{d}t}\Big(T_{\underline{a}}\,\mathtt{P}^{\underline{a}} + M_{\underline{a}\,\underline{b}}\,\mathtt{J}^{\underline{a}\,\underline{b}}\Big) := \Big\{H\big[N,N^a\big],H\big[M,M^a\big]\Big\}\Big|_{\Gamma}\ . \quad (34)$$

Evaluating the right hand side of (34) by using (27), for the time independence of the physical quantities above we obtain the following list:

$$\dot{\mathtt{P}}^{\mathbf{0}} = 0 \quad \text{iff} \quad \beta_{\mathbf{k}}(t) = 0\,, \quad (35)$$

$$\dot{\mathtt{P}}^{\mathbf{i}} = 0 \quad \text{iff} \quad \beta_{\mathbf{k}}(t) = 0\,,\ \rho_{\mathbf{ki}}(t) = 0\,, \quad (36)$$

$$\dot{\mathtt{J}}^{\mathbf{ij}} = 0 \quad \text{iff} \quad \beta_{\mathbf{k}}(t) = 0\,,\ \rho_{\mathbf{ki}}(t) = 0\,,\ \tau_{\mathbf{i}}(t) = 0\,, \quad (37)$$

$$\dot{\mathtt{J}}^{\mathbf{i0}} = 0 \quad \text{iff} \quad \beta_{\mathbf{k}}(t) = 0\,,\ \rho_{\mathbf{ki}}(t) = 0,,\ \tau_{\mathbf{i}}(t) = 0,\ \tau(t) = 0\ . \quad (38)$$

Therefore, *the total ADM energy-momentum* $\mathtt{P}^{\underline{a}}$ *and the relativistic angular momentum* $\mathtt{J}^{\underline{a}\,\underline{b}}$*, built from the spatial Regge–Teitelboim angular momentum and the Beig–Ó Murchadha centre-of-mass, are conserved only with respect to gauge evolutions, i.e. when* $(N, N^a) \in \mathcal{G}$.

4.4 Three Difficulties

In Subsects. 2.5 and 3.2 we found that the lapse and the shift that ensure the conservation of the total energy-momentum and (relativistic) angular momentum of the matter fields may even be asymptotically linearly diverging, i.e. they may be any element of $\mathcal{A}$. In the light of this result it is quite surprising that the analogous gravitational quantities are conserved only with respect to considerably more restricted lapses and shifts: These must tend to zero at infinity, and, in particular, the Beig–Ó Murchadha centre-of-mass is not conserved even with respect to time evolution that is a pure asymptotic time translation at infinity. Thus we raise the question of whether the total energy-momentum and (relativistic) angular momentum introduced above are really the "ultimate" expressions, or whether there is a slightly different definition for them with better conservation properties. We expect that these total quantities must be conserved at least with respect to pure asymptotic time translations.

However, there is a second difficulty too. Although we noted in Subsects. 3.2 and 4.3 that the structure of the allowed lapses and shifts are only *roughly* similar to that of the time and space projections of the Killing fields in Minkowski spacetime, respectively, in Subsect. 4.3 we swept this observation under the rug, and we considered the elements of ${}_0\mathcal{A}$ as the lapse and shift parts of the generators of the 'asymptotic symmetries' of the spacetime. Nevertheless, strictly speaking, neither the elements of $\mathcal{A}$ nor of ${}_0\mathcal{A}$ can be identified with the generators of the asymptotic symmetries of the spacetime. Indeed, while the elements of $\mathcal{A}$ have arbitrary time dependence and the elements of ${}_0\mathcal{A}$ are completely time independent, the components of the Killing vectors of the Minkowski spacetime have a very specific, namely *linear* time dependence. In particular, the familiar boost Killing vectors of the Minkowski spacetime cannot be recovered, neither from $\mathcal{A}$ nor from ${}_0\mathcal{A}$, in the weak field approximation.

The third difficulty is that while the centre-of-mass of the matter fields in Minkowski spacetime depends on the Cartesian time coordinate, the Beig–Ó Murchadha centre-of-mass is completely time independent. But the time dependence of the centre-of-mass was needed to prove not only its conservation, but also its correct transformation properties in the spacetime. Although the relativistic angular momentum built from the spatial angular momentum and the Beig–Ó Murchadha centre-of-mass transforms in the correct way *in the phase space*, this does not imply its correct transformation *in the spacetime*.

In the rest of this contribution we try to resolve these three problems by showing first how the "correct" time dependence of the lapse functions can be obtained. Since these resolutions grew up from the need to have a systematic spacetime interpretation of the results and the analysis of Beig and Ó Murchadha, we go back to spacetime.

5 The Asymptotic Spacetime Killing Vectors

5.1 The 3 + 1 Form of the Lie Brackets and the Killing Operators

Let Σ be a smooth spacelike hypersurface with future pointing timelike unit normal t^a and induced metric q_{ab}. Let K^a and $\bar{K}^a$ be two arbitrary vector fields on M, and let their $3+1$ decomposition on Σ be $K^a = Mt^a + M^a$ and $\bar{K}^a = \bar{M}t^a + \bar{M}^a$. Then the $3+1$ decomposition of their Lie bracket with respect to Σ can be written as

$$\begin{aligned}\left[K,\bar{K}\right]^a &= \left(t^a t^b + 2q^{ab}\right)\left(M\nabla_{(b}\bar{K}_{c)} - \bar{M}\nabla_{(b}K_{c)}\right)t^c \\ &\quad + t^a\left(\text{Ł}_{\mathbf{M}}\bar{M} - \text{Ł}_{\bar{\mathbf{M}}}M\right) + \left(\left[M,\bar{M}\right]^a - \left(MD^a\bar{M} - \bar{M}D^aM\right)\right). \quad (39)\end{aligned}$$

Observe that the first two terms on the right are the time-time and the time-space projections of the spacetime Killing operators acting on K^a and $\bar{K}^a$. The third term on the right is precisely the combination of the lapse and shift parts of K^a and $\bar{K}^a$ that appeared as the new lapse in the calculation of the Poisson bracket of two Beig–Ó Murchadha Hamiltonians (27). Similarly, the last term is built from M, M^a, $\bar{M}$ and $\bar{M}^a$ precisely in the same way as the new shift from the old lapses and shifts in (27). Thus one can expect that the Lie bracket of spacetime vector fields plays some role in the Poisson algebra of the Beig–Ó Murchadha Hamiltonians. Parts of the Killing operator are vanishing in some sense. Therefore it is worth decomposing the Killing operator in the $3+1$ way as well.

Although the space-space projection of the Killing operator can be expressed by three dimensional quantities defined with respect to Σ, the time-time and the time-space projections can be done only if we have not only a single spacelike hypersurface, but a whole foliation and a notion of "time flow" ξ^a as well. Thus we fix the vector field ξ^a, which will be represented by a lapse and a shift according to $\xi^a = Nt^a + N^a$. If $\dot{X}^a$ denotes the projection of the Lie derivative of the *spatial* X^a along ξ^a, then the full $3+1$ decomposition of $\nabla^{(a}K^{b)}$ is

$$Nt_c t_d \nabla^{(c}K^{d)} = \dot{M} + \text{Ł}_{\mathbf{M}}N - \text{Ł}_{\mathbf{N}}M\,, \quad (40)$$

$$2NP^a_c t_d \nabla^{(c}K^{d)} = \dot{M}^a + \left(ND^aM - MD^aN\right) - \left[N,M\right]^a, \quad (41)$$

$$P^a_c P^b_d \nabla^{(c}K^{d)} = D^{(a}M^{b)} + M\chi^{ab}\,. \quad (42)$$

Recall that precisely these projections appeared in (23). Furthermore, apart from the dot-derivatives, the right hand side of (40) and (41) are precisely the special combinations of the lapses and shifts that already appeared in (27). Equations (39–41) will be our key equations.

5.2 The Asymptotic Killing Vectors

In Subsect. 3.2 we introduced $\mathcal{A}$ as the space of the allowed, most general lapse-shift pairs compatible with the boundary conditions via the evolution equations. Thus in this picture $\mathcal{A}$ is the space of the allowed spacetime coordinate systems based on a single, fixed asymptotically flat spacelike hypersurface Σ. Two elements of $\mathcal{A}$, say (M, M^a) and (M', M'^a), determine two different foliations of the spacetime, and the corresponding unit timelike normals, t^a and t'^a, are different.

However, we can look at the space $\mathcal{A}$ from a slightly different perspective too. Let us fix a vector field ξ^a, which determines a foliation of the spacetime that is based on the single asymptotically flat Σ. Let t^a be the future pointing unit timelike normal of the leaves of this foliation, and let the lapse and the shift parts of ξ^a be chosen to be allowed: $(N, N^a) \in \mathcal{A}$, where $Nt^a + N^a = \xi^a$. Then for any $(M, M^a) \in \mathcal{A}$ define the spacetime vector field $K^a := Mt^a + M^a$. Note that we use the *same* t^a to define K^a for all (M, M^a). Thus the role of ξ^a is to provide a differential topological background to build spacetime vector fields from the pairs (M, M^a). The space of such spacetime vector fields will be denoted by $\mathcal{A}_\xi$, and let $\mathcal{G}_\xi$ be its subspace whose elements are constructed using $\mathcal{G}$ instead of $\mathcal{A}$.

Next observe that the space-space projection of the Killing operator (42) acting on any vector field $K^a \in \mathcal{A}_\xi$ is vanishing asymptotically at least as $O(r^{-k})$, and if this fall-off is actually $O(r^{-k})$ then the leading term has even parity. However, its time-time and time-space projections can still be arbitrary. This motivates us how to define the asymptotic Killing vectors: The vector field $K^a \in \mathcal{A}_\xi$ will be called an *asymptotic Killing vector with respect to* ξ^a if $t^c t^d \nabla_{(c} K_{d)}$ and $P^c_a t^d \nabla_{(c} K_{d)}$ are also vanishing asymptotically at least as $O(r^{-k})$, and if this fall-off is actually $O(r^{-k})$ then the leading terms have even parity. We can introduce a slightly stronger notion: $K^a \in \mathcal{A}_\xi$ will be called a *strongly asymptotic Killing vector with respect to* ξ^a if $t^c t^d \nabla_{(c} K_{d)} = 0$ and $P^c_a t^d \nabla_{(c} K_{d)} = 0$, i.e. when the right side of (40) and (41) is vanishing not only asymptotically, but pointwise as well. Note that although the Killing equation has only the trivial solution in a general spacetime, the asymptotic Killing and the strong asymptotic Killing equations can always be solved among the vector fields $K^a \in \mathcal{A}_\xi$.

Indeed, $t^c t^d \nabla_{(c} K_{d)} = O(r^{-k})$ and $P^c_a t^d \nabla_{(c} K_{d)} = O(r^{-k})$ are *not* partial differential equations, they are only ordinary differential equations for the time dependence of M and M^a. In particular, if the asymptotic structure of the lapse N and the shift N^a is given by (21) and (22), respectively, and the asymptotic structure of $(M, M^a) \in \mathcal{A}$ is

$$M\big(t,x^{\mathbf{k}}\big)=2x^{\mathbf{k}}B_{\mathbf{k}}\big(t\big)+T\big(t\big)+r^G\mu^{(G)}\Big(t,\frac{x^{\mathbf{k}}}{r}\Big)+o^\infty\big(r^G\big)\,, \tag{43}$$

$$M_{\mathbf{i}}\big(t,x^{\mathbf{k}}\big)=2x^{\mathbf{k}}R_{\mathbf{ki}}\big(t\big)+T_{\mathbf{i}}\big(t\big)+r^H\mu^{(H)}_{\mathbf{i}}\Big(t,\frac{x^{\mathbf{k}}}{r}\Big)+o^\infty\big(r^H\big)\,, \tag{44}$$

where $G,H\leq(1-k)$, then both the asymptotic and the strong asymptotic Killing equations give the ordinary differential equations

$$\dot{B}_{\mathbf{i}}=-2\Big(R_{\mathbf{ij}}\beta^{\mathbf{j}}-\rho_{\mathbf{ij}}B^{\mathbf{j}}\Big)\,, \tag{45}$$

$$\dot{R}_{\mathbf{ij}}=2\Big(B_{\mathbf{i}}\beta_{\mathbf{j}}-\beta_{\mathbf{i}}B_{\mathbf{j}}\Big)-2\Big(R_{\mathbf{ik}}\rho^{\mathbf{k}}{}_{\mathbf{j}}-\rho_{\mathbf{ik}}R^{\mathbf{k}}{}_{\mathbf{j}}\Big)\,, \tag{46}$$

and if $k\geq 1$, we also have

$$\dot{T}=-2\Big(T_{\mathbf{i}}\beta^{\mathbf{i}}-\tau_{\mathbf{i}}B^{\mathbf{i}}\Big)\,, \tag{47}$$

$$\dot{T}_{\mathbf{i}}=2\Big(T\beta_{\mathbf{i}}-\tau B_{\mathbf{i}}\Big)-2\Big(T^{\mathbf{j}}\rho_{\mathbf{ji}}-\tau^{\mathbf{j}}R_{\mathbf{ji}}\Big)\,. \tag{48}$$

(45)–(48) is a system of ordinary differential equations for $B_{\mathbf{i}}(t)$, $R_{\mathbf{ij}}(t)$, $T_{\mathbf{i}}(t)$ and $T(t)$. For given $\beta_{\mathbf{i}}(t)$, $\rho_{\mathbf{ij}}(t)$, $\tau_{\mathbf{i}}(t)$ and $\tau(t)$ this can always be solved, and the solution depends on six, and if $k\geq 1$ then on ten constants of integration. Here raising and lowering of the boldface Roman indices are defined by the spatial projection of the constant Minkowski metric: $\eta_{\mathbf{ij}}=-\delta_{\mathbf{ij}}$. Thus the role of the asymptotic Killing equations is that they restrict the unspecified time dependence of M and M^a. In particular,

- if $\rho_{\mathbf{ij}}=0$, $\beta_{\mathbf{i}}=0$, $\tau_{\mathbf{i}}=0$ and $\tau=0$, i.e. if ξ^a is a pure gauge generator $(N,N^a)\in\mathcal{G}$, then the spacetime coordinate system that ξ^a defines is asymptotically *collapsing*. Then the solution of (45)-(48) is that $B_{\mathbf{i}}$, $R_{\mathbf{ij}}$, $T_{\mathbf{i}}$ and T are all constant;
- if $\rho_{\mathbf{ij}}=0$, $\beta_{\mathbf{i}}=0$, $\tau_{\mathbf{i}}=0$ and $\tau=1$, i.e. if ξ^a is a pure asymptotic time translation, then the corresponding coordinate system is an *asymptotically Cartesian coordinate system*. Then $B_{\mathbf{i}}$, $R_{\mathbf{ij}}$ and T are constant but $T_{\mathbf{i}}(t)=T_{\mathbf{i}}-2tB_{\mathbf{i}}$ for some constants $T_{\mathbf{i}}$;
- if $\rho_{\mathbf{ij}}=0$, $\tau=0$, $\beta_{\mathbf{i}}=$ const. with $\beta_{\mathbf{i}}\beta_{\mathbf{j}}\delta^{\mathbf{ij}}=1$ and $\tau_{\mathbf{i}}(t)=-2t\beta_{\mathbf{i}}$, then the corresponding coordinates form an *asymptotically Rindler coordinate system*. Then the solution of (45)–(48) is considerably more complicated:

$$
\begin{aligned}
B_{\mathbf{i}}(t) &= -\beta_{\mathbf{i}}\beta^{\mathbf{k}}B_{\mathbf{k}} + \Pi_{\mathbf{i}}^{\mathbf{k}}B_{\mathbf{k}}\cosh(2t) - R_{\mathbf{ik}}\beta^{\mathbf{k}}\sinh(2t)\,,\\
R_{\mathbf{ij}}(t) &= \Pi_{\mathbf{i}}^{\mathbf{k}}\Pi_{\mathbf{j}}^{\mathbf{l}}R_{\mathbf{kl}}\\
&\quad + \big(\beta_{\mathbf{i}}R_{\mathbf{jk}} - \beta_{\mathbf{j}}R_{\mathbf{ik}}\big)\beta^{\mathbf{k}}\cosh(2t) - \big(\beta_{\mathbf{i}}\Pi_{\mathbf{j}}^{\mathbf{k}} - \beta_{\mathbf{j}}\Pi_{\mathbf{i}}^{\mathbf{k}}\big)B_{\mathbf{k}}\sinh(2t)\,,\\
T(t) &= \beta^{\mathbf{k}}B_{\mathbf{k}} + \big(T - \beta^{\mathbf{k}}B_{\mathbf{k}}\big)\cosh(2t) - \beta^{\mathbf{k}}T_{\mathbf{k}}\sinh(2t)\,,\\
T_{\mathbf{i}}(t) &= \Pi_{\mathbf{i}}^{\mathbf{k}}T_{\mathbf{k}} + R_{\mathbf{ik}}\beta^{\mathbf{k}} + 2t\beta_{\mathbf{i}}\beta^{\mathbf{k}}B_{\mathbf{k}}\\
&\quad - \Big(\beta_{\mathbf{i}}\beta^{\mathbf{k}}T_{\mathbf{k}}2t\Pi_{\mathbf{i}}^{\mathbf{k}}B_{\mathbf{k}} + R_{\mathbf{ik}}\beta^{\mathbf{k}}\Big)\cosh(2t)\\
&\quad + \Big(\beta_{\mathbf{i}}\big(T - \beta^{\mathbf{k}}B_{\mathbf{k}}\big) + \Pi_{\mathbf{i}}^{\mathbf{k}}B_{\mathbf{k}} + 2tR_{\mathbf{ik}}\beta^{\mathbf{k}}\Big)\sinh(2t)\,,
\end{aligned}
$$

where $\Pi_{\mathbf{i}}^{\mathbf{k}} := \delta_{\mathbf{i}}^{\mathbf{k}} + \beta_{\mathbf{i}}\beta^{\mathbf{k}}$ is the projection to the 2-plane orthogonal to $\beta_{\mathbf{i}}$ and T, $T_{\mathbf{i}}$, $B_{\mathbf{i}}$ and $R_{\mathbf{ij}}$ are constants.

Therefore, both the time independent generators of Beig and Ó Murchadha and the familiar Killing vectors of the Minkowski spacetime can be recovered as asymptotic Killing vectors by an appropriate choice for ξ^a, and the latter is connected with the asymptotically Cartesian coordinate system discussed in Subsect. 2.4.

The space of the asymptotic Killing vectors and of the strong asymptotic Killing vectors (with respect to ξ^a) will be denoted by $\mathcal{A}_\xi^K$ and $\mathcal{A}_\xi^0$, respectively, and obviously $\mathcal{A}_\xi^0 \subset \mathcal{A}_\xi^K \subset \mathcal{A}_\xi$.

5.3 The Algebra of Asymptotic Symmetries

Contrary to expectations, the space $\mathcal{A}_\xi$ does not close to a Lie algebra with respect to the spacetime Lie bracket. To see this, it is enough the consider the t^a component of the Lie bracket given by (39) and take into account that $\text{Ł}_{\mathbf{M}}\bar{M} - \text{Ł}_{\bar{\mathbf{M}}}M$ has the form of an allowed lapse for any $(M, M^a), (\bar{M}, \bar{M}^a) \in \mathcal{A}$, while the leading term in $t^a t^b(M\nabla_{(a}\bar{K}_{b)} - \bar{M}\nabla_{(a}K_{b)})$ has the form $N^{-1}x^{\mathbf{i}}x^{\mathbf{j}}$, which deviates from the structure of the allowed lapses.

If K^a and $\bar{K}^a$ are any two asymptotic Killing vectors then by (39) their Lie bracket, $\hat{K}^a := [K, \bar{K}]^a$, belongs to $\mathcal{A}_\xi$. Furthermore, the (time dependent) parameters in its asymptotic expansion according to (43) and (44), $\hat{B}_{\mathbf{i}}$ and $\hat{R}_{\mathbf{ij}}$, and if $k \geq 1$ then $\hat{T}_{\mathbf{i}}$ and $\hat{T}$ also, are built from those of K^a and $\bar{K}^a$ as

$$\hat{B}_{\mathbf{i}} = 2\big(R_{\mathbf{ij}}\bar{B}^{\mathbf{j}} - \bar{R}_{\mathbf{ij}}B^{\mathbf{j}}\big)\,, \tag{49}$$

$$\hat{R}_{\mathbf{ij}} = 2\big(R_{\mathbf{ik}}\bar{R}^{\mathbf{k}}{}_{\mathbf{j}} - \bar{R}_{\mathbf{ik}}R^{\mathbf{k}}{}_{\mathbf{j}} + \bar{B}_{\mathbf{i}}B_{\mathbf{j}} - B_{\mathbf{i}}\bar{B}_{\mathbf{j}}\big)\,, \tag{50}$$

$$\hat{T}_{\mathbf{i}} = 2\big(T^{\mathbf{j}}\bar{R}_{\mathbf{ji}} - \bar{T}^{\mathbf{j}}R_{\mathbf{ji}} + \bar{T}B_{\mathbf{i}} - T\bar{B}_{\mathbf{i}}\big)\,, \tag{51}$$

$$\hat{T} = 2\big(T_{\mathbf{i}}\bar{B}^{\mathbf{i}} - \bar{T}_{\mathbf{i}}B^{\mathbf{i}}\big)\,. \tag{52}$$

Now it is a direct calculation to show that $\hat{B}_{\mathbf{i}}$, $\hat{R}_{\mathbf{ij}}$, $\hat{T}_{\mathbf{i}}$ and $\hat{T}$ satisfy (45)-(48). Thus the leading, and if $k \geq 1$ then the leading two terms in $[K, \bar{K}]^a$

satisfy even the strong asymptotic Killing equations. However, in general $[K,\bar{K}]^a$ does not satisfy the asymptotic Killing equations even if both K^a and $\bar{K}^a$ are strong asymptotic Killing vectors. To see this we should calculate the projections $P^c_a P^d_b \text{Ł}_{[\mathbf{K},\bar{\mathbf{K}}]} g_{cd}$, $P^c_a t^d \text{Ł}_{[\mathbf{K},\bar{\mathbf{K}}]} g_{cd}$ and $t^c t^d \text{Ł}_{[\mathbf{K},\bar{\mathbf{K}}]} g_{cd}$. Using the differential geometric identity $L_{[\mathbf{K},\bar{\mathbf{K}}]} = \text{Ł}_{\mathbf{K}} \text{Ł}_{\bar{\mathbf{K}}} - \text{Ł}_{\bar{\mathbf{K}}} \text{Ł}_{\mathbf{K}}$, it is a straightforward calculation to show that $P^c_a t^d \text{Ł}_{[\mathbf{K},\bar{\mathbf{K}}]} g_{cd}$ is not of order $O(r^{-k})$ for general $K^a, \bar{K}^a \in \mathcal{A}^K_\xi$, and it is not zero for general $K^a, \bar{K}^a \in \mathcal{A}^0_\xi$. Therefore, neither $\mathcal{A}^K_\xi$ nor $\mathcal{A}^0_\xi$ close to a Lie algebra. Nevertheless, by the fact that $\hat{B}_{\mathbf{i}}$, $\hat{R}_{\mathbf{ij}}$, $\hat{T}_{\mathbf{i}}$ and $\hat{T}$ satisfy (45)–(48) the Lie bracket of any two asymptotic Killing vectors deviates from an asymptotic Killing field only by an element of $\mathcal{G}_\xi$. This observation makes it possible to introduce a natural Lie algebra structure on the quotient vector spaces $\mathcal{A}^K_\xi/\mathcal{G}^K_\xi$ and $\mathcal{A}^0_\xi/\mathcal{G}^0_\xi$, where $\mathcal{G}^K_\xi := \mathcal{G}_\xi \cap \mathcal{A}^K_\xi$ and $\mathcal{G}^0_\xi := \mathcal{G}_\xi \cap \mathcal{A}^0_\xi$. These quotient spaces are spanned by the (special time dependent) parameters $B_{\mathbf{i}}$ and $R_{\mathbf{ij}}$, and if $k \geq 1$ then also by $T_{\mathbf{i}}$ and T. Hence they are isomorphic to each other and their dimension is six for $k < 1$ and ten for $k \geq 1$. The Lie multiplication of them is given by (49)–(52), and it is easy to see that this Lie algebra is the Lorentz Lie algebra for $k < 1$ and the Poincaré algebra for $k \geq 1$. Therefore, *the structure of the Lie algebra* $\mathcal{A}^K_\xi/\mathcal{G}^K_\xi$ *is connected with the fall-off rate of the metric: for slow fall-off it is only the Lorentz Lie algebra, and the displacements of the origin of the coordinate system emerge as asymptotic symmetries only for* $1/r$ *or faster fall-off.*

6 Beig–Ó Murchadha Hamiltonians with Asymptotic Spacetime Killing Vectors

In this section we return to the discussion of the properties of the Beig–Ó Murchadha Hamiltonian, but instead of the elements of the time independent $(M, M^a) \in {}_0\mathcal{A}$ we parameterize them by the asymptotic Killing vectors.

Thus let us fix ξ^a, and define $H[K^a] := H[M, M^a]$ for any $K^a := Mt^a + M^a \in \mathcal{A}^K_\xi$. Then by (39) the Lie multiplication law (27) in the Poisson algebra of the Beig–Ó Murchadha Hamiltonians can be written in the remarkably simple form

$$\left\{H\left[K^a\right], H\left[\bar{K}^a\right]\right\} = \begin{cases} -H\left[\left[K,\bar{K}\right]^a\right] + \text{constraints} & \text{for } K^a, \bar{K}^a \in \mathcal{A}^K_\xi\,, \\ -H\left[\left[K,\bar{K}\right]^a\right] & \text{for } K^a, \bar{K}^a \in \mathcal{A}^0_\xi\,. \end{cases} \tag{53}$$

Therefore, *apart from constraints, the Beig–Ó Murchadha Hamiltonian preserves the spacetime Lie bracket of the asymptotic spacetime Killing vectors, and it preserves the spacetime Lie bracket of the strong asymptotic spacetime Killing vectors.*

The second issue that we consider is the conservation of the Hamiltonian. Thus let $(M, M^a) \in \mathcal{A}$, and calculate the *total* time derivative of $H[M, M^a]$, where the time evolution is generated by $\xi^a = Nt^a + N^a$. Then

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t} H\left[M, M^a\right] &= H\big[\dot{M}, \dot{M}^a\big] + \Big\{H\big[N, N^a\big], H\big[M, M^a\big]\Big\} \\ &= H\big[\dot{M} + M^e D_e N - N^e D_e M, \dot{M}^a + N D^a M - M D^a N - [N, M]^a\big] \\ &= \begin{cases} \text{constraints for } Mt^a + M^a \in \mathcal{A}^K_\xi, \\ 0 \qquad\qquad\quad \text{for } Mt^a + M^a \in \mathcal{A}^0_\xi\,. \end{cases} \end{aligned} \tag{54}$$

Here we used (27), and, in the last step, the definition of the asymptotic Killing and the strong asymptotic Killing vectors. Thus, *the Beig–Ó Murchadha Hamiltonian is constant (constant modulo constraints) with respect to the time evolution defined by* ξ^a *if* $K^a = Mt^a + M^a$ *is strongly asymptotic Killing (asymptotic Killing) with respect to* ξ^a.

7 Physical Quantities from the Beig–Ó Murchadha Hamiltonians with Asymptotic Spacetime Killing Vectors

7.1 The General Definition of the Physical Quantities

Independently of the details of the canonical analysis of the vacuum Einstein theory, we can consider the Beig–Ó Murchadha Hamiltonian as a functional of the initial data on an asymptotically flat spacelike hypersurface even in the presence of matter fields and even if the fall-off rate of the metric is assumed only to be positive. Thus for any $(M, M^a) \in \mathcal{A}$ let us define

$$\begin{aligned}\mathtt{Q}\big[M, M^a\big] &:= H\big[M, M^a\big]|_\Gamma + \mathtt{Q}^m\big[M, M^a\big] \\ &= -\frac{1}{2\kappa}\int_\Sigma D_a \Big(M q^{ab} q^{cd}\big({}_0D_c q_{bd} - {}_0D_b q_{cd}\big) \\ &\qquad + \big({}_0D_b M\big) q^{ab} q^{cd}\big(q_{cd} - {}_0q_{cd}\big) \\ &\qquad - \big({}_0D_c M\big) q^{ab} q^{cd}\big(q_{bd} - {}_0q_{bd}\big) \\ &\qquad - 2M_b\big(\chi^{ba} - \chi q^{ba}\big)\Big)\sqrt{|q|}\mathrm{d}^3x\,. \end{aligned} \tag{55}$$

Apparently, for zero $B_{\mathbf{i}}$ and $R_{\mathbf{ij}}$ but non-zero T or $T_{\mathbf{i}}$ this expression is finite only if $k \geq 1$. However, as it was pointed out in [6], [7], [8] in the vacuum case, $\mathtt{Q}[M, M^a]$ is finite even if $k > 1/2$ *and the fall-off rate* G *and* H *in (43)–(44) satisfies the stronger restriction* $G, H \leq -k$: Relaxing the fall off for the matter fields analogously, the right hand side can be written as the sum of a finite and a would-be divergent term, but the latter in fact vanishes by the constraint parts of the field equations. (Of course, in this case the

energy-momentum of the matter fields is *not* finite.) Similarly, apparently $\mathtt{Q}[M, M^a]$ can be finite for non-zero $B_{\mathbf{i}}$ and $R_{\mathbf{ij}}$ only for $k \geq 2$, but, as an analogous analysis shows [14], the slowest possible fall-off rate ensuring the finiteness of (55) is in fact $k \geq 1$.

7.2 Total Energy, Momentum, Angular Momentum and Centre-of-Mass

Next let us restrict $K^a := Mt^a + M^a$ to be an asymptotic Killing vector and introduce the notation $\mathtt{Q}[K^a] := \mathtt{Q}[M, M^a]$. Then since $\mathcal{A}^K_\xi/\mathcal{G}^K_\xi \approx \mathcal{A}^0_\xi/\mathcal{G}^0_\xi$ is coordinatized by the integration *constants* $B_{\mathbf{i}}$ and $R_{\mathbf{ij}}$, and for $k \geq 1$ by $T_{\mathbf{i}}$ and T too, $\mathtt{Q}[K^a]$ is a linear expression of them:

$$\mathtt{Q}[K^a] = T\mathtt{P}^{\mathbf{0}} + T_{\mathbf{i}}\mathtt{P}^{\mathbf{i}} + R_{\mathbf{ij}}\mathtt{J}^{\mathbf{ij}} + 2B_{\mathbf{i}}\mathtt{J}^{\mathbf{i0}}. \tag{56}$$

This defines the total energy, linear momentum, spatial angular momentum and centre-of-mass, respectively. However, these quantities depend on the choice of the vector field ξ^a. In particular,

- if ξ^a is chosen to be a pure gauge generator, then we recover the ADM energy $\mathtt{P}^{\mathbf{0}}_{ADM}$, the ADM linear momentum $\mathtt{P}^{\mathbf{i}}_{ADM}$, the Regge–Teitelboim spatial angular momentum $\mathtt{J}^{\mathbf{ij}}_{RT}$ and the Beig–Ó Murchadha centre-of-mass $\mathtt{J}^{\mathbf{i0}}_{BOM}$, respectively;
- if ξ^a is a pure asymptotic time translation, then the energy, linear momentum and spatial angular momentum coincide with the ADM energy and linear momentum and the Regge–Teitelboim angular momentum, but the centre-of-mass deviates slightly from the Beig–Ó Murchadha centre-of-mass; it is $\mathtt{J}^{\mathbf{i0}} = \mathtt{J}^{\mathbf{i0}}_{BOM} - t\mathtt{P}^{\mathbf{i}}_{ADM}$;
- if ξ^a defines an asymptotically Rindler coordinate system, then the energy, linear momentum, spatial angular momentum and centre-of-mass will be complicated time dependent combinations of the ADM energy and linear momentum, the Regge–Teitelboim angular momentum and the Beig–Ó Murchadha centre-of-mass:

$$\begin{aligned}
\mathtt{P}^{\mathbf{0}} &= \mathtt{P}^{\mathbf{0}}_{ADM}\cosh(2t) + \beta_{\mathbf{k}}\mathtt{P}^{\mathbf{k}}_{ADM}\sinh(2t),\\
\mathtt{P}^{\mathbf{i}} &= \Pi^{\mathbf{i}}_{\mathbf{k}}\mathtt{P}^{\mathbf{k}}_{ADM} - \beta^{\mathbf{i}}\big(\mathtt{P}^{\mathbf{0}}_{ADM}\sinh(2t) + \beta_{\mathbf{k}}\mathtt{P}^{\mathbf{k}}_{ADM}\cosh(2t)\big),\\
\mathtt{J}^{\mathbf{ij}} &= \Pi^{\mathbf{i}}_{\mathbf{k}}\Pi^{\mathbf{j}}_{\mathbf{l}}\mathtt{J}^{\mathbf{kl}}_{RT} + 2\beta^{[\mathbf{i}}\mathtt{J}^{\mathbf{j}]\mathbf{k}}_{RT}\beta_{\mathbf{k}}\cosh(2t) + 2\beta^{[\mathbf{i}}\mathtt{J}^{\mathbf{j}]\mathbf{0}}_{BOM}\sinh(2t)\\
&\quad - \beta^{[\mathbf{i}}\mathtt{P}^{\mathbf{j}]}_{ADM}\big(1 - \cosh(2t) + 2t\sinh(2t)\big),\\
\mathtt{J}^{\mathbf{i0}} &= -\beta^{\mathbf{i}}\beta_{\mathbf{k}}\mathtt{J}^{\mathbf{k0}}_{BOM} + \Pi^{\mathbf{i}}_{\mathbf{k}}\big(\mathtt{J}^{\mathbf{k0}}_{BOM}\cosh(2t) + \mathtt{J}^{\mathbf{kl}}_{RT}\beta_{\mathbf{l}}\sinh(2t)\big)\\
&\quad + \frac{1}{2}\mathtt{P}^{\mathbf{i}}_{ADM}\sinh(2t) + t\big(\beta^{\mathbf{i}}\beta_{\mathbf{k}} - \Pi^{\mathbf{i}}_{\mathbf{k}}\cosh(2t)\big)\mathtt{P}^{\mathbf{i}}_{ADM}\\
&\quad + \frac{1}{2}\beta^{\mathbf{i}}\big(1 - \cosh(2t)\big)\mathtt{P}^{\mathbf{0}}_{ADM}\,.
\end{aligned}$$

Thus the definition of the physical quantities, defined by the value of the Beig–Ó Murchadha Hamiltonian parameterized by the asymptotic spacetime Killing vectors, do depend on the vector field ξ^a that we used to define the asymptotic Killing vectors. Hence we should have a selection rule for ξ^a. Based on the discussions in Subsect. 2.4, such a selection rule could be the requirement that the spacetime coordinate system determined by ξ^a be asymptotically Cartesian. Our suggestion is to take such a ξ^a. In fact, this choice should be justified by the properties of the corresponding physical quantities.

The analysis of Subsect. 4.3 to clarify the transformation properties of this total energy, linear momentum, spatial angular momentum and centre-of-mass can be repeated. It is easy to see that they have exactly the same transformation properties *in the phase space* that the quantities defined in Subsect. 4.3 had: They form a Lorentzian 4-vector $\mathtt{P}^{\underline{a}}$ and an anti-symmetric tensor $\mathtt{J}^{\underline{a}\,\underline{b}}$, and transform according to the Poincaré transformation. However, defining the Cartesian spacetime coordinates by $x^{\underline{a}} := (t, x^{\mathbf{i}})$, we can consider the transformation of $\mathtt{P}^{\underline{a}}$ and $\mathtt{J}^{\underline{a}\,\underline{b}}$ under the Poincaré transformation of the Cartesian coordinates, $x^{\underline{a}} \mapsto x^{\underline{b}}\,\Lambda_{\underline{b}}{}^{\underline{a}} + C^{\underline{a}}$, *in the spacetime* too. Using the explicit form of M and M^a in terms of the spacetime Cartesian coordinates and the defining equation (56), it is a straightforward calculation to show that $\mathtt{P}^{\underline{a}}$ and $\mathtt{J}^{\underline{a}\,\underline{b}}$ transform just in the correct way. It might be worth noting that the special *linear* time dependence of the centre-of-mass is needed to derive the correct transformation properties. In fact, the relativistic angular momentum tensor built from the Regge–Teitelboim angular momentum and the Beig–Ó Murchadha centre-of-mass does *not* transform in the expected way under Poincaré transformations of the Cartesian coordinates $x^{\underline{a}}$ *in the spacetime*.

Next let us consider again a general ξ^a, and calculate the *total* time derivative of $\mathtt{Q}[K^a]$ with respect to ξ^a. Now the coefficients in the asymptotic spacetime Killing vectors K^a have explicit time dependence. Using the evolution equations of Subsect. 3.2, we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathtt{Q}\big[K^a\big] &= \mathtt{Q}\big[\dot{M} + M^e D_e N - N^e D_e M, \dot{M}^a + N D^a M - M D^a N - [N, M]^a\big] \\ &= 0 \end{aligned} \tag{57}$$

for any $K^a \in \mathcal{A}^K_\xi$. Therefore, *the energy-momentum and angular momentum, defined by* $\mathtt{Q}[K^a]$ *with the vector fields* K^a *that are asymptotic Killing with respect to* ξ^a*, are conserved in time provided the time evolution is defined by the same* ξ^a. Thus, just as in Subsects. 2.5 and 3.2, the vector field ξ^a defining the time evolution is required only to be an allowed time axis, but the generators K^a for the physical quantities do depend on ξ^a. In particular, both the conservation (35)–(38) of the time independent quantities with respect to gauge evolutions in Subsect. 4.3 and the conservation of the energy-momentum and relativistic angular momentum defined in the present subsection with respect to pure asymptotic time translations are special cases of (57).

7.3 Translations for Slow Fall-Off Metrics

In Subsects. 5.2 and 5.3 we saw that the asymptotic translations emerge as genuine asymptotic symmetries only for $1/r$ or faster fall-off, while for slow fall-off they are lost in the sea of the "generators of gauge evolutions" and the genuine asymptotic symmetries are only the asymptotic rotations and boosts. On the other hand, by the results of Subsect. 7.1, for slow, $1/r^k$, $1/2 < k < 1$, fall-off we can define energy-momentum but not relativistic angular momentum. The aim of the present subsection is to resolve this apparent contradiction by showing what the translations in the slow fall-off case might be.

The key observation is that $\mathtt{Q}[M, M^a]$ can be finite for the slow fall-off metrics provided the structure of M and M^a is

$$M\left(t, x^{\mathbf{k}}\right) = T\left(t\right) + r^K \mu^{(K)}\left(t, \frac{x^{\mathbf{k}}}{r}\right) + o^\infty\left(r^K\right), \tag{58}$$

$$M_{\mathbf{i}}\left(t, x^{\mathbf{k}}\right) = T_{\mathbf{i}}\left(t\right) + r^L \mu^{(L)}_{\mathbf{i}}\left(t, \frac{x^{\mathbf{k}}}{r}\right) + o^\infty\left(r^L\right), \tag{59}$$

where $K, L \leq -k$, i.e. the $x^{\mathbf{k}}$-dependent parts of M and M^a tend to zero as r^{-k} rather than diverging as $r^{(1-k)}$ as in (43) and (44). This motivates us to consider for some $q \leq (1-k)$ the spacetime vector fields $K^a = Mt^a + M^a$ whose asymptotic structure is given by (58)–(59) and $K, L \leq q$. We say that they have q-fast fall-off. In general, these vector fields do not form a Lie algebra.

Next consider the space ${}_q\mathcal{T}^K_\xi$ of such vector fields which are asymptotic Killing vectors too: Let the $t^a t^b \nabla_{(a} K_{b)}$ and $P^a_c t^b \nabla_{(a} K_{b)}$ parts of the Killing operator acting on them tend to zero at least as $O(r^{q-1})$. Then the Lie bracket $[K, \bar{K}]^a$ of $K^a \in {}_q\mathcal{T}^K_\xi$ and $\bar{K}^a \in \mathcal{A}^K_\xi$ contains terms of order r^{-k}. Thus the Lie bracket operation preserves the index q of the space ${}_q\mathcal{T}^K_\xi$ and the components of $[K, \bar{K}]^a$ have the structure (58)-(59) only if $q \geq -k$. The quotient ${}_q\mathcal{T}^K_\xi / {}_q\mathcal{T}^K_\xi \cap \mathcal{G}^K_\xi$ is isomorphic to $\mathbb{R}^4$ and inherits a commutative Lie algebra structure from $\mathcal{A}^K_\xi / \mathcal{G}^K_\xi$. Equations (47) and (48) show that $T(t)$ and $T_{\mathbf{i}}(t)$ are in fact constant for ξ^a generating e.g. an asymptotically collapsing or asymptotically Cartesian coordinate system. (If ξ^a generates an asymptotically Rindler coordinate system, then they still depend on time as $T(t) = T\cosh(2t) + T^*\sinh(2t)$ and $T_{\mathbf{i}}(t) = T_{\mathbf{i}} + \beta_{\mathbf{i}}(T^*\cosh(2t) + T\sinh(2t))$ for constants T, T^* and $T_{\mathbf{i}}$ satisfying $T_{\mathbf{i}}\beta^{\mathbf{i}} = 0$.) Thus ${}_q\mathcal{T}^K_\xi$ may be interpreted as the space of the "q-fast fall-off asymptotic translations" in $\mathcal{A}^K_\xi$ even if $k \in (0,1)$, provided $-k \leq q \leq (1-k)$. On the other hand, by the results of Subsect. 7.1 the translations yielding finite energy-momentum can be the elements of ${}_q\mathcal{T}^K_\xi$ for any $q \leq (1-k)$ if $k \geq 1$, but for $0 < k < 1$ only those "q-fast fall-off" translations yield finite energy-momentum for which $q \leq -k$. Therefore, the space of the fast fall-off translations yielding finite energy-momentum is precisely ${}_{-k}\mathcal{T}^K_\xi$.

8 Summary

The present investigation grew from the need to give a systematic *space-time* interpretation of the results and the main points of the analysis of canonical vacuum general relativity by Beig and Ó Murchadha. However, while the centre-of-mass components of the relativistic angular momentum of matter fields in Minkowski spacetime depend linearly on time, the Beig–Ó Murchadha centre-of-mass expression for asymptotically flat spacetimes is completely time independent. As a consequence of this the Beig–Ó Murchadha centre-of-mass is conserved only with respect to "gauge evolutions", and although it transforms in the correct way in the phase space, it does not in the spacetime.

To find the correct time dependence we suggest to parameterize the Beig–Ó Murchadha Hamiltonian by the lapse and shift parts of appropriately defined asymptotic Killing vector fields. A natural Lie algebra structure can be introduced on the quotient of the space of the asymptotic Killing fields and the subspace of "gauge generators", and we showed that this Lie algebra is only the Lorentz Lie algebra for slow fall-off, but it is the Poincaré algebra for $1/r$ or faster fall-off metrics.

We define the total energy-momentum and relativistic angular momentum by the value on the constraint surface of the Beig–Ó Murchadha Hamiltonian parameterized by the *asymptotic translation or rotation-boost Killing vectors*. This definition is completely analogous to that of the (quasi-local or total) energy-momentum and angular momentum of matter fields using the Killing vectors of the Minkowski spacetime. The energy-momentum obtained in this way is just the standard ADM energy-momentum and the spatial angular momentum is that of Regge and Teitelboim. However, the centre-of-mass deviates from that of Beig and Ó Murchadha by a term, which is the linear momentum times the coordinate time. This centre-of-mass has the correct transformation properties, known for the matter fields in flat spacetime, both in the phase space and in the spacetime with respect to asymptotic Poincaré transformations, and it is conserved if the time evolution is generated by asymptotic time translations.

Acknowledgements

I am grateful to Robert Beig for his valuable remarks on angular momentum at spatial infinity. Special thanks to the organizers of the 319 WE–Heraeus-Seminar on Mathematical Relativity and the Wilhelm and Else Heraeus Foundation itself for the hospitality at Bad Honnef. This work was partially supported by the Hungarian Scientific Research Fund grant OTKA T042531.

References

1. J. Frauendiener: *Conformal Infinity*. Living Rev. Relativity **7** (2004), 1 `http://www.livingreviews.org/lrr-2004-1` 158
2. H. Friedrich: Gravitational fields near space-like and null infinity. J. Geom. Phys. **24**, 83–163 (1998) 158
3. R. Arnowitt, S. Deser, C.W. Misner: The dynamics of general relativity. In: *Gravitation: An Introduction to Current Research*, ed by L. Witten (Wiley, New York 1962) Ch. 7, pp 227–265 158, 171
4. T. Regge, C. Teitelboim: Role of surface integrals in the Hamiltonian formulation of general relativity. Ann. Phys. (N.Y.) **88**, 286–318 (1974) 158, 170, 171
5. R. Beig, N. Ó Murchadha: The Poincaré group as the symmetry group of canonical general relativity. Ann. Phys. (N.Y.) **174**, 463–498 (1987) 158, 159, 168, 170
6. P.T. Chruściel: Boundary conditions at spacelike infinity from a Hamiltonian point of view. In: *Topological and Global Structure of Spacetime*, NATO Adv.Sci. Inst. Ser. B Phys. vol 138, (Plenum, New York 1986) pp. 49–59 158, 179
7. R. Bartnik: The mass of an asymptotically flat manifold. Commun. Pure. Appl. Math. **39**, 661–693 (1986) 158, 179
8. N. Ó Murchadha: Total energy-momentum in general relativity. J. Math. Phys. **27**, 2111–2128 (1986) 158, 179
9. R. Schoen, S.-T. Yau: Proof of the positive mass theorem. II. Commun. Math. Phys. **79**, 231–260 (1981) 158
10. E. Witten: A new proof of the positive energy theorem. Commun. Math. Phys. **30** 381–402 (1981) 158
11. L.B. Szabados: *Quasi-Local Energy-Momentum and Angular Momentum in GR: A Review Article*. Living Rev. Relativity **7** (2004), 4 `http://www.livingreviews.org/lrr-2004-4` 158, 161
12. J.M. Nester, F.-H. Ho, C.-M. Chen: Quasilocal center-of-mass for teleparallel gravity. `gr-qc/0403101v1` 158
13. J.M. Nester, F.-F. Meng, C.-M. Chen: Quasilocal center-of-mass. `gr-qc/0403103v2` 158
14. L.B. Szabados: On the Poincaré structure of asymptotically flat spacetimes. Class. Quantum Grav. **20** 2627–2661 (2003) 158, 159, 163, 167, 180

Part III

Numerical Methods

Computer Simulation – a Tool for Mathematical Relativity – and Vice Versa

Beverly K. Berger

Physics Division, National Science Foundation, Arlington, VA 22207, USA
bberger@nsf.gov

Abstract. Examples from the study of spatially inhomogeneous cosmological space-times are given to illustrate the potential for synergy between mathematical analysis and computer simulation.

1 Introduction

In this contribution, I will discuss examples where computer simulations have provided insight into the properties of strong field gravity. Some of the research reported was done in collaboration with David Garfinkle, James Isenberg, Vincent Moncrief, and Marsha Weaver. For additional details and other examples see [9].

The singularity theorems due to Penrose, Hawking, and others [25] state that regular, generic initial data for reasonable matter will evolve to yield a pathological behavior if the gravitational field becomes sufficiently strong. However, these theorems do not describe the nature of the resulting pathology and examples of several types are known. The relevance or lack thereof to the everyday world are described by Penrose's Cosmic Censorship Conjectures (CCCs) [37]. These state that, generically, singularities will be hidden inside the horizons of black holes. This means that naked singularities do not occur in nature – i.e., singularities that do occur cannot influence the space-time exterior to the event horizon of the black hole that contains them. A stronger form of the CCC states that time-like singularities will not occur generically even inside a horizon. This means that an observer will only detect a singularity by hitting it. No time-like observer can detect signals from a space-like singularity in his future. Consider, for example, the Penrose conformal diagrams for the Schwarzschild and Reissner-Nordström space-times. The space-like singularity in the Schwarzschild solution obeys the CCC because the final singularity is within the black hole event horizon, is space-like, and lies to the future of any observer falling into the black hole. An observer falling into this singularity cannot receive any information from it until falling into it. In contrast, the singularities in the Reissner-Nordström singularity are time-like (even though they are within an event horizon). Signals from the

B.K. Berger: *Computer Simulation – a Tool for Mathematical Relativity – and Vice Versa*, Lect. Notes Phys. **692**, 187–203 (2006)
www.springerlink.com

singularities could reach (i.e. these singularities would be visible to) a timelike observer.

There are several motivations for the study of singularities. Singularities signal a breakdown of classical General Relativity. The nature of the breakdown may be relevant for quantum gravity since quantum gravity is often invoked as a way to avoid this breakdown of the classical theory. If CCC is false, naked singularities might exist and be important astrophysically. Finally, on route to a singularity, strong gravitational fields – interesting in their own right – will be encountered. As an arena to study the approach to the singularity and other features of strong field (nonlinear) gravity, we shall focus on spatially inhomogeneous cosmological space-times. Cosmological space-times do not necessarily describe the actual universe. They are solutions to Einstein's equations with "cosmological boundary conditions" in contrast to the asymptotically flat space-times that describe localized systems such as binary black holes. In this discussion, we shall consider only space-times with T^3 spatial topology. Clearly, however, the actual universe is spatially inhomogeneous. These inhomogeneities are generally neglected in cosmological studies. What role do the spatial inhomogeneities play in the nature of singularities and in cosmologies in general?

One way to study singularities is to search for a way to characterize the approach to the singularity. Long ago, Belinskii, Lifshitz, and Khalatnikov (BKL) [4, 3] conjectured that the approach to the singularity that arose in generic gravitational collapse had a very simple form as illustrated in Fig. 1. We can imagine evolving the PDEs that are Einstein's equations on a spacetime grid in the collapse direction. The BKL conjecture states that for any spatial point, sufficiently close to the singularity, the spatial derivatives that connect spatial points on the grid become dynamically unimportant and the evolution may be described as the solution to the ODEs of a spatially homogeneous cosmology. This means that each spatial point evolves eventually as a separate universe.

One may then ask if numerical simulation can be used to explore the nature of the approach to the singularity and the BKL conjecture. Here we shall describe just such a program as an example of the synergy between mathematical and numerical techniques to provide insight and understanding in the approach to singularities and other properties of cosmological space-times. Both mathematical and numerical methods have advantages and disadvantages. Mathematical techniques yield theorems that describe large classes of solutions at once. On the other hand, these techniques often gain their power by avoiding the need to obtain information about details of the solutions. Of course, this is the forte of numerical simulation. However, such simulations can study only one solution at a time so that it may be necessary to explore a large parameter space to understand classes of solutions.

The following examples show how numerical simulation can be used to make a strong statement in situations where one might naively expect this

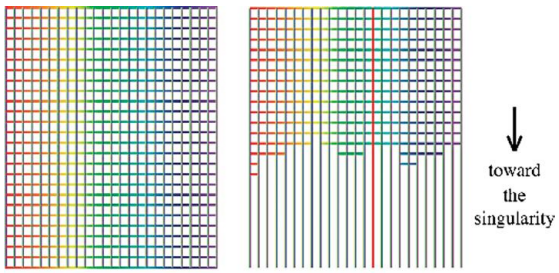

Fig. 1. Cartoon of the BKL conjecture. The grid at the left represents the space (*horizontal*) and time (*vertical downward*) evolution of the Einstein equation PDEs. BKL claim that sufficiently close to the singularity for any given spatial point, the PDEs may be replaced at that point by the ODEs of an appropriate spatially homogeneous cosmology

tool to have little use. The first example is the simple ODE

$$\frac{d^2 f}{dx^2} + \frac{2}{x}\frac{df}{dx} + f = 0 \tag{1}$$

for a real function $f(x)$. The general solution is, of course,

$$f(x) = a\,\frac{\sin x}{x} + b\,\frac{\cos x}{x} \tag{2}$$

where a and b are arbitrary constants. Note that the first term on the right hand side of (2) has a finite limit as $x \to 0$ while the second term is infinite in that limit. A generic numerical solution to (1) will display the asymptotic behavior of $\cos x/x$ as $x \to 0$ since that term will be much larger than the $\sin x\,/\,x$ term in the limit. Thus, in the 2-parameter space of solutions described by the coefficients a and b, the solution $b = 0$ is a non-generic set of measure zero. To discover the existence and properties of the non-generic solution numerically, one may take advantage of the change in sign of the limit of $f(x)$ as b passes through zero. This is illustrated in Fig. 2. Start at a large value of x, say x_i and integrate (1) toward $x = 0$. Vary the initial data (e.g. hold $f(x_i)$ fixed and vary df/dx at x_i) to "zero in on" the initial data equivalent to $b = 0$. This technique is essentially that used to explore critical phenomena in general relativity. This simple example also illustrates how one can explore the properties of a singularity (the generic $f(0)$ is infinite) even though computers cannot proceed in the presence of singular values.

A second example illustrates the point that the presence of an attractor, i.e. a single asymptotic in time solution that arises from generic initial data, in effect allows computer simulations to describe large classes of solutions without the need to explore in detail the space of initial data. Figure 3 shows the superposition of three solutions of the Lorenz system of equations [33]

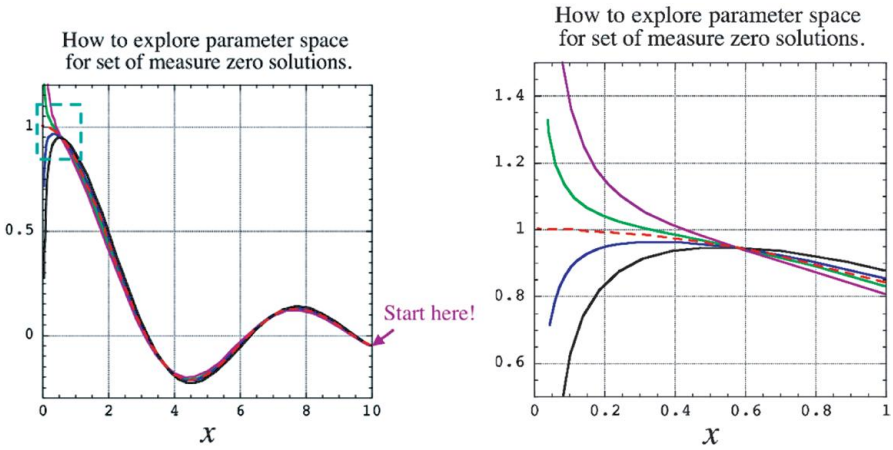

Fig. 2. Finding a set-of-measure-zero solution. The figure on the left displays numerical solutions of (1) starting from $x_0 = 10$ with different values of $\dot{x}_0$. The inset is shown as the figure on the right

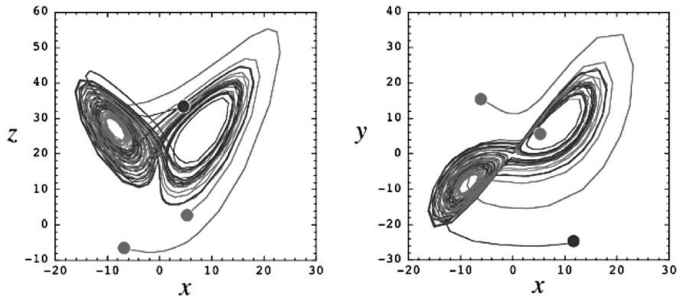

Fig. 3. The Lorenz attractor. The graphs show two projections of numerical solutions to (3) starting from 3 sets of initial conditions marked by the dots

$$\begin{aligned} \frac{dx}{dt} &= 10\,(y\,-\,x), \\ \frac{dy}{dt} &= 28\,x\,-\,y\,-\,x\,z, \\ \frac{dz}{dt} &= x\,y\,-\,\frac{8\,z}{3}\,, \end{aligned} \tag{3}$$

for x, y, and z functions of t. As can be seen, the location of the Lorenz attractor may be determined by starting from a large open set of initial points.

2 Mixmaster Dynamics and the BKL Conjecture

In most of the remainder of this paper, we shall discuss the behavior of cosmological space-times. We shall further specialize to vacuum cosmological space-times even though these are not relevant to the actual universe. However, the singularity structure in vacuum space-times illustrates the properties of strong field gravity and avoids complications due to singularities in matter, such as shock waves in fluids, that occur even in the absence of gravitational fields.

2.1 How Spatially Homogeneous Cosmologies Collapse

Large classes of spatially homogeneous cosmologies can be described in the "Hamiltonian formulation" (see [41]). We shall consider vacuum, anisotropic models described by three scale factors $R_x(t)$, $R_y(t)$, and $R_z(t)$ rather than the single $R(t)$ that describes the Friedmann–Robertson–Walker spatially homogeneous, isotropic models. We follow Misner's minisuperspace (MSS) notation [35] in replacing these scale factors by equivalent variables $\{\Omega, \beta_\pm\}$ denoting respectively the volume and anisotropic shears. For the models of interest here, Einstein's equations may be obtained from the variation of the Hamiltonian constraint which takes the form

$$2H^0 = -p_\Omega^2 + p_+^2 + p_-^2 + V(\Omega, \beta_+, \beta_-) \tag{4}$$

where $\{p_\Omega, p_\pm\}$ are canonically conjugate to $\{\Omega, \beta_\pm\}$ and V is a "potential" that arises from the spatial scalar curvature.[1]

The simplest vacuum cosmology is the Kasner model described by three anisotropic scale factors with power law dependence on comoving proper time t. They are described by the metric

$$ds^2 = -dt^2 + \sum_{i=1}^{3} t^{2p_i}\, dx_i^2 \tag{5}$$

where

$$\sum_{i=1}^{3} p_i^2 = 1 = \sum_{i=1}^{3} p_i \tag{6}$$

and the space-like hypersurfaces are flat. Note that (6) implies that, in collapse, one of the axes will (almost)[2] always expand although the spatial volume t will decrease. In terms of the MSS variables, the potential V vanishes yielding the Hamiltonian constraint

[1] Note that this Hamiltonian could be generalized by the addition of a number of arbitrary constants. Here we assume that these have been absorbed by rescaling the metric.

[2] There is an exceptional case for exponents {1, 0, 0 }. The term "(almost)" will be used throughout this Chapter to indicate the existence of set-of-measure-zero exceptional cases.

$$2H^0 = -p_\Omega^2 + p_+^2 + p_-^2 = 0. \tag{7}$$

Defining $v_\pm = p_\pm/p_\Omega$ allows the replacement of (7) by

$$v_+^2 + v_-^2 = 1 \tag{8}$$

with the solution (taking Ω to play the role of time)[3]

$$\beta_\pm = v_\pm \left|\Omega\right| . \tag{9}$$

In the MSS with axes $\{\Omega, \beta_\pm\}$, the Kasner solution is a straight line. As we shall demonstrate explicitly later, the role of the potential V in other spatially homogeneous models is to induce a change from one Kasner solution to another. In many classes of spatially homogeneous cosmologies, in the direction of collapse to the singularity, there is a final Kasner epoch. In spatially inhomogeneous models, there may be (in a sense to be clarified later) a final Kasner epoch at each spatial point. We refer to space-times with a final Kasner epoch in their approach to the singularity as asymptotically velocity term dominated (AVTD) [29].

If there is no final Kasner epoch, the approach to the singularity is said to exhibit local (in the spatially inhomogeneous case) Mixmaster dynamics (LMD). This phenomenon was first identified by BKL [4] and then independently by Misner [34] who coined the name.[4] The archetypical model in this class is the diagonal, vacuum Bianchi IX cosmology described (in MSS) by the metric [34]

$$ds^2 = -e^{3\Omega} d\tau^2 + e^{2\Omega} \left(e^{2\beta}\right)_{ij} d\sigma^i d\sigma^j \tag{10}$$

where the anisotropic metric components are obtained from exponentiation of $\beta = \mathrm{diag}(-2\beta_+, \beta_+ + \sqrt{3}\beta_-, \beta_+ - \sqrt{3}\beta_-)$, all variables depend only on the BKL time τ, and the spatial 1-forms σ^i satisfy the appropriate $SU(2)$ relationship for Bianchi IX. Einstein's equations may be found from the variation of (4) [34] for

$$V(\Omega, \beta_\pm) = e^{4\Omega - 8\beta_+} + e^{4\Omega + 4\beta_+ + 4\sqrt{3}\beta_-} + e^{4\Omega + 4\beta_- + 4\sqrt{3}\beta_-} + \cdots \tag{11}$$

where the ellipsis indicates terms that are (almost) always negligible. See, however, the discussion in e.g. [6]. While decades of numerical simulation indicated the plausibility of a never ending sequence of Kasner epochs in this model, proofs of various aspects of Mixmaster dynamics have appeared only recently [44, 39].

In the absence of the potential V (with $\sigma^i = dx^i$), the Kasner solution is obtained. Equation (4) with (11) for V defines the dynamics in minisuperspace (MSS). The Kasner solution represents the free particle in MSS. For the Kasner solution, (8) may be written as

[3] Some arbitrary constants have been ignored.

[4] Mixmaster is a 1950s era brand of kitchen appliance.

$$K \equiv v_+^2 + v_-^2 = 1 \,, \tag{12}$$

where $v_\pm = -p_\pm/p_\Omega$. The addition of the potential (11) causes (almost) every Kasner epoch to end in a bounce off one of the first three terms on the right hand side. After the bounce, the behavior is again described by (12) but with different Kasner parameters. As first discussed by BKL, every Kasner epoch can be identified by a single parameter u (related to the anisotropic collapse rates) such that the $n+1$st Kasner epoch is related to the nth one (in Mixmaster dynamics) through

$$u_{n+1} = \begin{cases} u_n - 1 & \text{for} \quad u_n \geq 2 \,, \\ \frac{1}{u_n - 1} & \text{for} \quad 1 \leq u_n \leq 2 \,. \end{cases} \tag{13}$$

The u-map (13) is an example of a "bounce law" found by using conservation of momentum between asymptotic Kasner solutions obtained from the Hamiltonian (4) with V replaced by the dominant exponential term in (11).

It is easy to determine the dominant term. Just assume a Kasner solution (9) and substitute $\beta_\pm$ into V from (11). As is discussed elsewhere [16], only one term will grow (almost always). (This dominance was exploited in a numerical scheme [12] to allow a Mixmaster evolution to be followed for hundreds of bounces with machine precision.) If the terms are separately monitored, it is seen that a peak in the largest term coincides with a change of Kasner epoch (i.e. a bounce). (See Fig. 1 in [16].) The elementary classical mechanics problem of scattering off an exponential potential can be used to relate ingoing and outgoing asymptotic momenta to yield bounce laws.

2.2 Do $U(1)$-Symmetric Cosmologies Exhibit LMD?

As an example of the role of numerical simulations in mathematical cosmology, we consider vacuum space-times on $T^3 \times R$ with a single spatial symmetry. These are described by the metric [10] (for a specific choice of lapse and shift)

$$ds^2 = e^{-2\varphi}\left[-e^{2\Lambda}\, d\tau^2 + e^{\Lambda}\, e_{ab}(x,z) d\xi^a d\xi^b\right] + e^{2\varphi}(d\xi^3 + \beta_a\, dx^a\, d\tau)^2 \tag{14}$$

where $a, b = 1,\ 2$ and $\varphi,\ \Lambda,\ x,\ z$, and β_a depend on spatial variables $\xi_1,\ \xi_2$, and BKL time τ. The metrics of this class have two "twist" fields β_a that satisfy a constraint $e^a\, \beta_a = 0$, for e^a canonically conjugate to β_a. The constraint is solved once and for all by replacing the two twist degrees of freedom by the single twist potential ω (and its conjugate momentum r). The properties of these models and their generalizations have been studied extensively by Moncrief [36]. The explicit form of e_{ab} is given in [36, 15] as is the discussion of a canonical transformation to replace the twists β_a with a single twist potential ω.

Since it is possible to express a Bianchi IX model on S^3 as a $U(1)$-symmetric model on S^3, one can identify LMD behavior in terms of the

$U(1)$ variables [18]. As is seen in Figs. 2 and 4 of [18], the $U(1)$ variable φ, related to the + polarization of the gravitational waves, tracks the largest scale factor of the collapsing Mixmaster model, while z decreases monotonically until a Mixmaster "era" ends.[5] While this comparison is made for S^3 spatial topology, BKL conjecture holds that the asymptotic dynamics of spatially inhomogeneous models should be that of separate universes at every spatial point. This local dynamics should be insensitive to global boundary conditions.

Numerical simulations of $U(1)$-symmetric cosmologies have yielded the expected behavior for φ [14].[6] Since ends of eras need not occur in the short simulations of [14], the predicted behavior of z was not seen there (see Fig. 6a of [14]). Tentative observation of the predicted behavior for z was obtained by exploration of the space of initial data. An example is given in Fig. 3 of [10]. More work needs to be done to confirm this.

Completely generic collapse has been studied by Garfinkle [22]. In that case, numerical stability was obtained by choosing the variables of Uggla et al. [42] which are better adapted to the dynamics. To analyze the simulations, spatial invariants are used to identify the local value of BKL parameter u at any given spatial point. Garfinkle was able to track the evolution of u at a representative spatial point through several bounces. The sequence of u-values agreed with those predicted by (13) to significantly better than 1%.

3 Mathematical-Numerical Synergy in Spatially Inhomogeneous Cosmologies

Spatially inhomogeneous cosmologies appear to have either AVTD or LMD approaches to their big crunch singularity. Table 1 shows references to numerical simulations and mathematical results that support this conjecture. Note that (except for spatially homogeneous space-times) the mathematical statements can be made only if the space-times are AVTD – although there have been recent suggestions for methods of attack in more general cases [42].

3.1 Gowdy Models as an Example

Vacuum, spatially inhomogeneous cosmologies with two spatial Killing vectors on $T^3 \times R$ without "twists" are described by the metric first given by Gowdy [23]:

[5] These era ending bounces yield $u_{n+1} = 1/(u_n - 1)$, are thus sensitive to initial conditions, and provide the source of the Mixmaster chaos (see e.g. [6]).

[6] These simulations are of rather poor quality since the Hamiltonian constraint is solved by hand and spatial smoothing is needed for stability. Confirmation of the results in [14] was given by Hern [26].

Table 1. Status of mathematical and numerical studies of singularities in collapsing cosmological space-times. All models are vacuum space-times except those where $+\phi$ indicates the presence of a scalar field

Model	Singularity Type	Mathematical	Numerical
Polarized Gowdy	AVTD	[29]	
Generic Gowdy	AVTD	[31]	[13, 11]
Polarized T^2	AVTD	[27]	
Magnetic Gowdy	LMD		[45]
Generic T^2	LMD		[17]
Polarized $U(1)$	AVTD	[28]	[15]
Generic $U(1)$	LMD		[14]
Generic $U(1)+\phi$	AVTD		[8]
Generic $+\phi$	AVTD	[1]	[21]
Generic	LMD		[22]

$$ds^2 = e^{(\lambda+\tau)/2}\left(-e^{-2\tau}\,d\tau^2 + d\theta^2\right) + e^{P-\tau}\left(d\sigma + Qd\delta\right)^2 + e^{-P-\tau}d\delta^2\,. \quad (15)$$

Einstein's equations consist of wave equations for the + polarization P and the $\times$ polarization Q, Hamiltonian and momentum constraints that are first order in the derivatives of the background λ and P, Q, and λ depend only on spatial variable θ and time τ. The wave equations may be obtained by variation of

$$2\mathcal{H} = \pi_P^2 + e^{-2P}\pi_Q^2 + e^{-2\tau}P_{,\theta}^{\,2} + e^{2(P-\tau)}Q_{,\theta}^{\,2} \quad (16)$$

where π_P, π_Q are canonically conjugate to P, Q. Note that $\mathcal{H}$ is $\neq 0$ and is not the Hamiltonian constraint. As $\tau \to \infty$, we may obtain the velocity term dominated (VTD) solution by neglecting spatial derivatives in the wave equations:

$$\begin{aligned} P \to v(\theta)\tau \quad &; \quad \pi_P \to v(\theta)\,; \\ Q \to Q_0(\theta) \quad &; \quad \pi_Q \to \pi_Q^0(\theta)\,. \end{aligned} \quad (17)$$

Substitution of (17) into the exponential terms in (16) shows that both exponential terms will decay if $0 < v(\theta) < 1$ but not otherwise [11]. If the VTD solution is consistent with the absence of the exponential potentials, we then conjecture that the approach to the singularity is AVTD. If $v(\theta)$ is <0 or >1, one of the exponential potentials will grow to cause a bounce. These bounces will eventually drive $v(\theta)$ into the AVTD range. The evolution of P at a representative spatial point in a numerical simulation validates this conjecture as may be seen in Fig. 2 of [16].

Spiky features in Gowdy spatial waveforms are understood to arise in the vicinity of "exceptional" points where the coefficient of one of the exponential potentials vanishes [11]. These features are now well understood analytically [38].

3.2 Expanding Gowdy Space-Times

The expanding polarized Gowdy cosmology was first studied analytically in [5] where a WKB analysis was used to describe the behavior as gravitational standing waves of decreasing amplitude in a background homogeneous space-time. (For a recent treatment see [30].) The wave amplitude was found to decay as $t^{-1/2}$ where $t = e^{-\tau}$. Similar results were found for other, related models including those with both wave polarizations [19, 20, 43, 32].

Notice that the VTD solution (17) is expressed in terms of a number of functions of space that are constant in time. These temporal constants are equivalent to true constants that describe the Kasner solution. Equivalently [13], these constants may be related to symmetries in the "target space" with metric

$$dS^2 = dP^2 + e^{2P} dQ^2 \ . \tag{18}$$

Since the behavior in the expanding direction does *not* have the *local* BKL character, the local constants are not relevant in this regime. However, certain combinations of these same constants have the property that their time derivative is a total spatial derivative. Thus, for expanding Gowdy models, we can explore the role of global constants[7]

$$\begin{aligned}
\bar{\alpha} &= \oint d\theta \left(2te^{2P} Q_{,t}\, Q - 2tP_{,t} \right) , \\
\bar{\beta} &= -\oint d\theta \left(te^{2P} Q_{,t} \right) , \\
\bar{\gamma} &= \oint d\theta \left(-tQ_{,t}\, Q + te^{2P} Q_{,t}\, Q^2 + 2tP_{,t}\, Q \right) \ .
\end{aligned} \tag{19}$$

Although these constants had been known for a long time [24], Ringström realized only recently that use of the constants to characterize the solutions to the wave equations for P and Q could yield some surprises [40]. For the Kasner model (written in the variables of (15)), the constants α, β, and γ defined as the integrands of (19) satisfy

$$\zeta_{Kasner} \equiv \frac{\alpha^2}{4} + \beta\gamma \ = \ t^2 \left(\dot{P}^2 + e^{2P} \dot{Q}^2 \right) \ \geq \ 0 \tag{20}$$

where the over-dot is d/dt in the spatially homogeneous model. However, because the average (i.e. the spatial integral over the circle) of a nonlinear function is not, in general, equal to the (same) nonlinear function of the averages, condition (20) need not hold when spatial dependence is allowed.

For

$$\bar{\zeta}_{Gowdy} \equiv \frac{\bar{\alpha}^2}{4} + \bar{\beta}\,\bar{\gamma} \geq 0 \ , \tag{21}$$

[7] As an aside we note that the integrands of these global constants can vary wildly with θ so that the quality of preservation of the integrals during a simulation is a good code test.

the behavior is as previously described of decaying waves in a spatially homogeneous background cosmology.

Polarized Gowdy Models As a first example, consider polarized ($Q = 0 = \pi_Q$) Gowdy models [5]. Einstein's equations are

$$0 = P_{,tt} + \frac{1}{t} P_{,t} - P_{,\theta\theta} \ , \tag{22}$$

$$0 = \lambda_{,t} - t \left(P_{,t}^2 + P_{,\theta}^2 \right) \ . \tag{23}$$

The wave equation for P has an explicit solution in terms of products of the form $\mathcal{Z}_0(nt) \cos(n\theta + \phi_n)$ where n is the integer mode number, $\mathcal{Z}_\nu(x)$ is an arbitrary Bessel function of order ν, and ϕ_n is an arbitrary constant phase. Since all zero-order Bessel functions have the large argument asymptotic expansion

$$\mathcal{Z}_0(nt) \approx \frac{\cos(nt + \xi_n)}{\sqrt{t}} \tag{24}$$

for ξ_n an arbitrary phase constant, the general solution to the wave equation (22) behaves asymptotically for $t \to \infty$ as

$$P(\theta, t) \approx P_0 + \zeta \ln t + \frac{1}{\sqrt{t}} S(\theta, t) \tag{25}$$

where $S(\theta, t)$ is periodic in θ, is $\mathcal{O}(1)$ in powers of t, and the leading term in $S_{,t}(\theta, t)$ is $\mathcal{O}(1)$ in powers of t. With this asymptotic form, (23) becomes

$$\lambda_{,t} \approx t \left(\frac{\zeta^2}{t^2} + \frac{S_{,t}^2}{t} + \frac{S_{,\theta}^2}{t} + 2 \frac{\zeta S_{,t}}{t^{3/2}} \right) \ . \tag{26}$$

If this were a Kasner model ($S = 0$), we would find

$$\lambda_{Kasner} = \lambda_0 + \zeta^2 \ln t \ . \tag{27}$$

However, the wave "energy" dominates in the Gowdy model to give

$$\bar{\lambda}_{Gowdy} \approx \overline{(S_{,t}^2 + S_{,\theta}^2)}\, t \tag{28}$$

where the over-bar is used to denote spatial averaging. See [5] for a discussion of the resultant background space-time. The "usual" understanding of expanding Gowdy models is that $\bar{P}$ and $\bar{Q}$ approach their Kasner forms (plus decaying waves) but $\bar{\lambda}$ evolves linearly rather than logarithmically in t.

Generic Gowdy Models Ringström recognized that, for Kasner, the Einstein equations for $P(t)$ and $Q(t)$ could be written in terms of the constants α, β, and γ so that three of the four constants in the solution would thus be determined. Earlier Moncrief and I had discussed using numerical simulations to find the point in the target space representing the asymptotic spatial averages of P and Q. (See [7].)

If $\bar{P}$ and $\bar{Q}$ are to have the asymptotic Kasner form, then the spatially averaged wave equations should approach

$$\begin{aligned} t\,\dot{\bar{P}} &= \bar{\beta}\,\bar{Q} - \frac{\bar{\alpha}}{2}\,, \\ t\,e^{\bar{P}}\,\dot{\bar{Q}} &= \bar{\beta}\,e^{-\bar{P}}, \\ t\,e^{\bar{P}}\,\dot{\bar{Q}} &= e^{\bar{P}}\,(\bar{\gamma} + \bar{\alpha}\,\bar{Q} - \bar{\beta}\,\bar{Q}^2) \end{aligned} \tag{29}$$

which are equivalent to the wave equations for P and Q obtained from the variation of (16) in the absence of spatial dependence.

However, Ringström proved [40] that $\bar{\zeta}$ defined by (21) could be negative. In that case, the solutions can be expected to have the asymptotic form

$$\begin{aligned} e^{\bar{P}} &= \frac{|\bar{\beta}|}{\sqrt{-\bar{\zeta}}}\left[c_1 - c_2\,\cos\left(\sqrt{-\bar{\zeta}}\,\ln\frac{t}{t_0}\right)\right], \\ \bar{Q} &= \frac{\bar{\alpha}}{2\bar{\beta}} + \frac{\sqrt{-\bar{\zeta}}}{|\bar{\beta}|}\,\frac{\left[c_2\,\sin\left(\sqrt{-\bar{\zeta}}\,\ln\frac{t}{t_0}\right)\right]}{\left[c_1 - c_2\,\cos\left(\sqrt{-\bar{\zeta}}\,\ln\frac{t}{t_0}\right)\right]} \end{aligned} \tag{30}$$

as was then seen by numerical simulations (see Fig. 4). Figure 5 illustrates the behavior of both sides of (29b) from a typical simulation (rewritten as $t\dot{\bar{Q}} = \bar{\beta}\,e^{-2\bar{P}}$). Clearly, as $t \to 0$, the difference decays for $\bar{\zeta} > 0$ but not for $\bar{\zeta} < 0$. Further details of this feature of expanding Gowdy space-times will be published elsewhere.

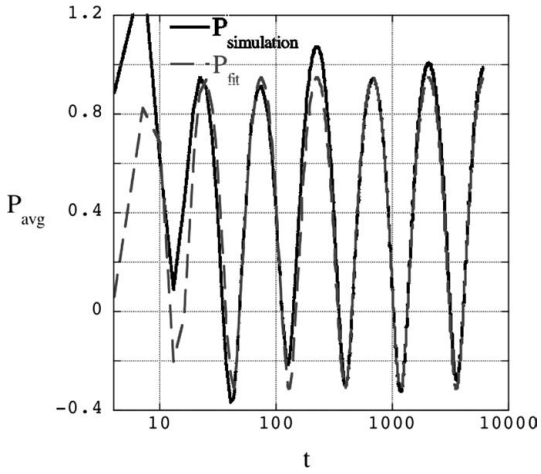

Fig. 4. Comparison of $\bar{P}$ from (30) (called P_{fit}) with the results of a simulation (called $P_{simulation}$) that starts from initial data with $\bar{\zeta} < 0$

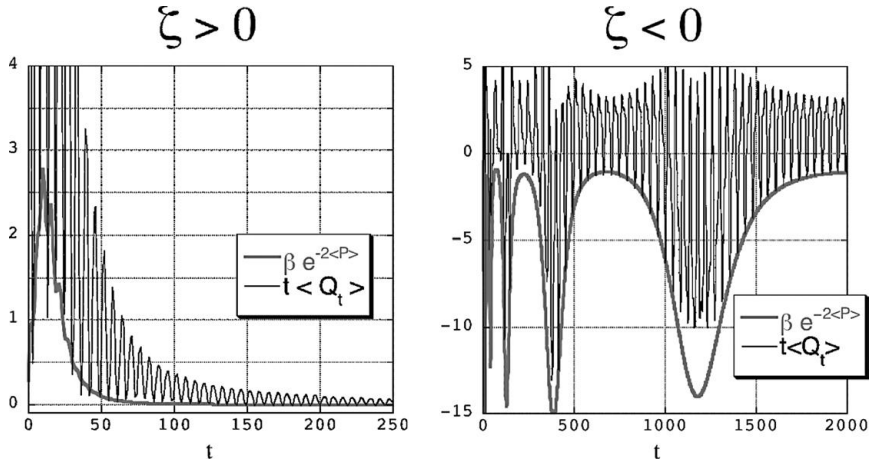

Fig. 5. The difference between the function of the averages and the averages of the function $t\dot{Q} = \beta\, e^{-2P}$ for either sign of $\bar{\zeta}$ (shown as ζ in the figure)

4 General T^2-Symmetric Space-Times as a "Laboratory" for Strong Field Gravity

As has long been recognized [23], Gowdy models on $T^3 \times R$ are obtained from the most general T^2-symmetric models by setting the twists (off-diagonal metric components of the form $g_{\theta x_i}$ where x_i is a symmetry direction) equal to zero. We have shown elsewhere [10] that, without loss of generality, these more general models may be described by

$$\begin{aligned} ds^2 = &- e^{(\lambda-3\tau)/2}\, d\tau^2 \,+\, e^{(\lambda+\mu+\tau)/2}\, d\theta^2 \\ &+ e^{P-\tau}\left\{dx \,+\, Q\, d\delta \,+\, \left[\int^{\tau} (Q\Theta) \,-\, Q\int^{\tau}\Theta\right] d\theta\right\}^2 \\ &+ e^{-P-\tau}\left[d\delta \,-\, \left(\int^{\tau}\Theta\right) d\theta\right]^2 \end{aligned} \tag{31}$$

where $\Theta = \kappa\, e^{\mu/4}\, e^{(\lambda+2P+3\tau)/2}$ for P, Q, and λ the Gowdy variables depending on τ, θ, $\pi_\lambda = e^{\mu/4}$ where π_λ is canonically conjugate to λ, and κ is the twist constant (see [17]). Einstein's equations may be found from the variation of [17]

$$\begin{aligned} H = &\frac{\pi_P^2}{4\pi_\lambda} \,+\, \frac{e^{-2\tau}\, P_{,\theta}^2}{4\pi_\lambda} \,+\, \frac{e^{-2P}\,\pi_Q^2}{4\pi_\lambda} \\ &+ \frac{e^{2(P-\tau)}\, Q_\theta^2}{4\pi_\lambda} \,+\, \kappa^2 \pi_\lambda\, e^{(\lambda+2P+3\tau)/2}\,. \end{aligned} \tag{32}$$

Just as in the Gowdy models, scattering off the separate exponential potential terms yields bounce laws. Unlike the Gowdy model, there is no regime for the

VTD solution that is consistent with all the potentials decaying. Thus, this class of models is expected to exhibit LMD (see the discussion in [17]). Support for this conjecture is found in the numerical verification of the derived bounce laws (see Table 3 and Figs. 5–8 in [17]).

In the expanding direction, these models are also interesting although only preliminary results have been obtained (in collaboration with J. Isenberg). For simplicity of the discussion here, we shall restrict attention to the polarized case. The equations become

$$P_{,tt} + \frac{1}{t}P_{,t} - \frac{P_{,\theta\theta}}{4\pi_\lambda^2} + \frac{\kappa^2 e^{\lambda/2+P}}{2t^{7/2}}(1 + tP_{,t}) + \frac{P_{,\theta}\,\pi_{\lambda,\theta}}{4\pi_\lambda^3} = 0\,, \tag{33}$$

$$\lambda_{,t} - t\left(P_{,t}^2 + \frac{P_{,\theta}^2}{4\pi_\lambda^2}\right) + \frac{\kappa^2 e^{\lambda/2+P}}{t^{5/2}} = 0\,, \tag{34}$$

$$\pi_{\lambda,t} - \frac{\kappa^2 e^{\lambda/2+P}}{t^{5/2}} = 0\,. \tag{35}$$

These equations differ from the polarized Gowdy ones (22)–(23) due to the twist terms. These appear to play the role in (33) of damping out most possible solutions to force the system toward a particular attractor.

A consistent analysis is as follows: Solve (35) by assuming $\bar{\pi}_\lambda \sim t^{-1/2}$. Then the fourth term on the left hand side of (33) provides damping unless $\bar{P} \sim -\ln t$. If $\bar{\lambda} \sim 5\ln t$, the exponential term on the left hand side of (34) will cancel the constant term in $\overline{\lambda_{,t}}$ that was found in the Gowdy case. Logarithmic time dependence of $\bar{\lambda}$ rather than the linear time dependence of the Gowdy case thus becomes the consistent solution. However, the coefficient 5 is inconsistent unless the average wave "energy"

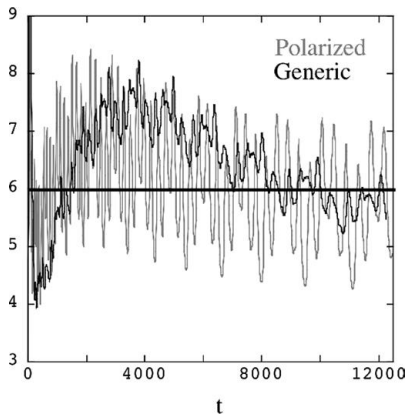

Fig. 6. Graph of $t\mathcal{E}$ from (36) vs t in a computer simulation of polarized and generic expanding T^2 symmetric cosmologies

$$\mathcal{E} \approx \oint d\theta\, t \left(P_{,t}^2 + \frac{P_{,\theta}^2}{4\pi_\lambda^2} \right) \sim \frac{6}{t} \,. \tag{36}$$

Surprisingly, this apparently bizarre attractor arises generically in simulations of these models. An example is shown in Fig. 6. Addition of $Q \neq 0$ does not change this result qualitatively although the deviations of the numerical simulation from the hypothesized attractor seem larger. Further exploration is in progress.

5 Conclusions

Numerical simulation has proven to be a valuable tool for the study of strong field gravity. Many of the mathematical studies of generic Gowdy collapse could proceed with confidence after the behavior of the solutions had been revealed numerically as may be seen in Table 1. Several controversies over Mixmaster dynamics in homogeneous space-times were clarified and resolved by numerical simulation.

Challenges remain, however. Mathematical techniques needed to study collapsing space-times with LMD are not known (see however the claims in [42] and a recent preprint [2]). Numerical simulation has provided insight into the behavior of these models.

Acknowledgements

This work was supported by the National Science Foundation.

References

1. L. Andersson, A.D. Rendall: Quiescent cosmological singularity. Commun. Math. Phys. **218**, 479–511 (2001) 195
2. L. Andersson, H. van Elst, W.C. Lim, C. Uggla: Asymptotic silence of generic cosmological singularities. Phys. Rev. Lett. **94**, 051101 (2005) 201
3. V.A. Belinskii, I.M. Khalatnikov, E.M. Lifshitz: A general solution of the Einstein equations with a time singularity. Adv. Phys. **31**, 639–667 (1982) 188
4. V.A. Belinskii, E.M. Lifshitz, I.M. Khalatnikov: Oscillatory approach to the singular point in relativistic cosmology. Sov. Phys. Usp. **13** 745–765 (1971) 188, 192
5. B.K. Berger: Quantum graviton creation in a model universe. Ann. Phys. (N.Y.) **83**, 458 (1974) 196, 197
6. B.K. Berger: Comments on the computation of Liapunov exponents for the Mixmaster universe. Gen. Rel. Grav. **23**, 1385 (1991) 192, 194
7. B.K. Berger: Asymptotic behavior of a class of expanding Gowdy spacetimes. `gr-qc/0207035` 197

8. B.K. Berger: Influence of scalar fields on the approach to the singularity in spatially inhomogeneous cosmologies. Phys. Rev. D **61**, 023508 (2000) 195
9. B.K. Berger: *Numerical approaches to spacetime singularities*. Living Rev. Relativity **5** (2002), 1. `http://www.livingreviews.org/lrr-2002-1` 187
10. B.K. Berger: Hunting local Mixmaster dynamics in spatially inhomogeneous cosmologies. Class. Quantum Grav. **21**, S81–S96 (2004) 193, 194, 199
11. B.K. Berger, D. Garfinkle: Phenomenology of the Gowdy model on $T^3 \times R$. Phys. Rev. D **57**, 4767 (1998) 195
12. B.K. Berger, D. Garfinkle, E. Strasser: New algorithm for Mixmaster dynamics. Class. Quantum Grav. **14** L29–L36 (1996) 193
13. B.K. Berger, V. Moncrief: Numerical investigation of cosmological singularities. Phys. Rev. D **48**, 4676 (1993) 195, 196
14. B.K. Berger, V. Moncrief: Evidence for an oscillatory singularity in generic $U(1)$ symmetric cosmologies on $T^3 \times R$. Phys. Rev. D **58** 064023 (1998) 194, 195
15. B.K. Berger, V. Moncrief: Numerical evidence for velocity dominated singularities in polarized $U(1)$ symmetric cosmologies. Phys. Rev. D **57**, 7235 (1998) 193, 195
16. B.K. Berger, D. Garfinkle, J. Isenberg, V. Moncrief, M. Weaver: The singularity in generic gravitational collapse is spacelike, local, and oscillatory. Mod. Phys. Lett. A **13**, 1565–1573 (1998) 193, 195
17. B.K. Berger, J. Isenberg, M. Weaver: Oscillatory approach to the singularity in vacuum spacetimes with T^2 isometry. Phys. Rev. D **62**, 123501 (2000) 195, 199, 200
18. B.K. Berger, V. Moncrief: Signature for local Mixmaster dynamics in $U(1)$ symmetric cosmologies. Phys. Rev. D **62**, 123501 (2000) 194
19. M. Carmeli, A. Feinstein: Inhomogeneous cosmologies: The cosmic peeling-off property of gravity. Int. J. Theor. Phys. **24** 1009 (1985) 196
20. A. Feinstein: Late-time behavior of primordial gravitational waves in expanding universe. Gen. Rel. Grav. **20**, 183 (1988) 196
21. D. Garfinkle: Harmonic coordinate method for simulating generic singularities. Phys. Rev. D **65**, 044029 (2002) 195
22. D. Garfinkle: Numerical simulations of generic singularities. Phys. Rev. Lett. **93**, 161101 (2004) 194, 195
23. R.H. Gowdy: Gravitational waves in closed universes. Phys. Rev. Lett. **27**, 826 (1971) 194, 199
24. B. Grubišić, V. Moncrief: Asymptotic behavior of the $T^3 \times R$ Gowdy spacetimes. Phys. Rev. D **47**, 2371–2382 (1993) 196
25. S.W. Hawking, R. Penrose: The singularities of gravitational collapse and cosmology. Proc. Roy. Soc. Lond. A **314**, 529–548 (1970) 187
26. S.D. Hern: Numerical relativity and inhomogeneous cosmologies. PhD thesis, Cambridge University (2000) 194
27. J. Isenberg, S. Kichenassamy: Asymptotic behavior in polarized T^2-symmetric vacuum spacetimes. J. Math. Phys. **40**, 340–352 (1999) 195
28. J. Isenberg, V. Moncrief: Asymptotic behavior of polarized and half-polarized $U(1)$ symmetric vacuum spacetimes. Class. Quantum Grav. **19**, 5361–5386 (2002) 195
29. J.A. Isenberg, V. Moncrief: Asymptotic behavior of the gravitational field and the nature of singularities in Gowdy spacetimes. Ann. Phys. (N.Y.) **199** 84 (1990) 192, 195
30. T. Jurke: On future asymptotics of polarized Gowdy T^3-models. Class. Quantum Grav. **20**, 173–192 (2003) 196

31. S. Kichenassamy, A.D. Rendall: Analytic description of singularities in Gowdy spacetimes. Class. Quantum Grav. **15** 1339–1355 (1998) 195
32. K.E. Kunze: Asymptotic behavior of inhomogeneous string cosmologies. Class. Quantum Grav. **16**, 3795–3806 (1999) 196
33. E.N. Lorenz: Deterministic nonperiodic flow. J. Atmos. Sci. **20**, 130–141 (1963) 189
34. C.W. Misner: Mixmaster universe. Phys. Rev. Lett. **22** 1071–1074 (1969) 192
35. C.W. Misner: Minisuperspace. In: *Magic Without Magic*, ed by J. Klauder (Freeman, San Francisco 1972) 191
36. V. Moncrief: Reduction of Einstein's equations for vacuum space-times with spacelike $U(1)$ isometry groups. Ann. Phys. (N.Y.) **167** 118 (1986) 193
37. R. Penrose: Gravitational collapse: The role of general relativity. Riv. Nuov. Cim. **1** 252–276 (1969) 187
38. A.D. Rendall, M. Weaver: Manufacture of Gowdy spacetimes with spikes. Class. Quantum Grav. **18**, 2959–2976 (2001) 195
39. H. Ringström: Curvature blow up in Bianchi VIII and IX vacuum spacetimes. Class. Quantum Grav. **17**, 713–731 (2000) 192
40. H. Ringström: On a wave map equation arising in general relativity. Commun. Pure Appl. Math. **57**, 657–703 (2004) 196, 198
41. M.P. Ryan, L.C. Shepley: *Homogeneous Relativistic Cosmologies*. (Princeton University, Princeton 1975) 191
42. C. Uggla, H. van Elst, J. Wainwright: The past attractor in inhomogeneous cosmology. Phys. Rev. D **68**, 103502 (2003) 194, 201
43. E. Verdaguer: Soliton solutions in spacetimes with two spacelike Killing fields. Phys. Reports **229**, 1–80 (1993) 196
44. M. Weaver: Dynamics of magnetic Bianchi VI_0 cosmologies. Class. Quantum Grav. **17**, 421–434 (2000) 192
45. M. Weaver, J. Isenberg, B.K. Berger: Mixmaster behavior in inhomogeneous cosmological spacetimes. Phys. Rev. Lett. **80**, 2984–2987 (1998) 195

On Boundary Conditions for the Einstein Equations

Simonetta Frittelli[1] and Roberto Gómez[2]

[1] Department of Physics, Duquesne University Pittsburgh, PA 15282, USA
simo@mayu.physics.duq.edu
[2] Pittsburgh Supercomputing Center, 4400 Fifth Avenue, Pittsburgh, PA 15213, USA
gomez@psc.edu

Abstract. The use of the projection of the Einstein tensor normally to a timelike boundary as a set of boundary conditions for the initial-value problem of the vacuum Einstein equations is investigated within the setting of a particular first-order strongly hyperbolic formulation. It is found that the components of such a projection give rise to boundary conditions that are appropriate, in a certain sense, for the initial-value problem of the evolution equations and for the initial-value problem of the auxiliary system of propagation of the constraints, at the same time. It can be concluded that imposing such boundary conditions on the values of the fundamental variables of the initial-value problem guarantees the propagation of the constraints. This contribution presents a unified account of results that have recently appeared separately in the literature. The presentation is meant to be accessible to a broader readership.

1 Introduction

In considering the Cauchy problem of the Einstein equations $G_{ab} = 0$ arising by performing a 3+1 split of the metric and the Einstein tensor [16, 15], one is – arguably unnaturally – led away from the geometric origin and meaning of the equations as a whole in favor of a purely analytical set up where a three dimensional Riemannian metric is determined, from initial values, by a set of second-order time-dependent partial differential equations of a rather mystic nature. While such a set up for the Einstein equations may be quite appealing to the Newtonian (i.e., analytic) mind, it carries a powerful drawback: the loss of the manifest four-dimensional character of the Einstein equations.

From the point of view of analysis, the geometric character of the equations is irrelevant to the initial-value problem. For over two decades now, since the influential paper by York where what is generally known as the ADM formulation of general relativity is laid down [17], a trend has developed to systematically forget where the ADM equations come from in order to focus, instead, on how to build a spacetime with them, by the might of computer power. But computer might alone has proven insufficient to solve problems of direct astrophysical relevance such as the gravitational wave

S. Frittelli and R. Gómez: *On Boundary Conditions for the Einstein Equations*, Lect. Notes Phys. **692**, 205–222 (2006)
www.springerlink.com

emission from the collision of two black holes. Progressively by trial and error, the influence of several factors on the results of the numerical integration has been identified. One such factor is the persistent refusal of numerical solutions to verify the constraints at some finite time after the start of the simulation. Under the spell of the remarkable historical development of the Cauchy problem of general relativity – which teaches, among other things, that once the constraints have been imposed on the initial data, they will be satisfied by the solution of the evolution equations *automatically* – this recalcitrant attitude of the numerical solutions ended up being entirely attributed to defects of the numerical schemes. But the fundamental difference between the Cauchy problem of the Einstein equations and the computational solution of the initial-value problem passed largely unrecognized. The difference lies at the boundary of the initial data surface: the necessity for consistent boundary conditions. The computational problem is not equivalent to the Cauchy problem, but to the initial-boundary value problem of the Einstein equations. The largely unrecognized fact is that if there is a boundary, the constraints cannot be satisfied by the solution even if they are imposed on the initial data with *infinite precision,* unless proper care is taken of the boundaries, again, with infinite precision. This is not a numerical problem but an analytical one. In other words, even the best numerical solutions of difference schemes are only mirroring what their perfect role models of the continuous equations do: tend to violate the constraints.

As far as we are aware of, the computational relevance of the connection between the constraints and the boundaries was first pointed out by Stewart in [14], on the basis of the analysis of the initial-boundary-value problems of the evolution equations of the fundamental variables themselves and of the associated evolution equations of the constraints. Stewart found that there is a set of boundary conditions that are appropriate for both initial-boundary value problems *simultaneously.* The *form* of the boundary conditions in [14] is not particularly illuminating, but that is not essential to our point: in a sense, [14] was, to our knowledge, the first to write what one may loosely refer to as constraint-preserving boundary conditions for general relativity. But a geometrical interpretation of this connection between the constraints and the boundaries was not developed in [14].

We wish to demonstrate that this connection arises in a natural way out of the four-dimensional setting and is endowed of a geometrical character. In hindsight, the reason why this connection does not stand out to the Newtonian mind can be traced directly to the lack of a manifestly four-dimensional nature in the ADM equations.

We consider the projection of the Einstein tensor G_{ab} along the direction normal to a timelike hypersurface and pursue the question of the place that the vanishing of such components of the Einstein tensor may have within the well-known $3 + 1$ setting of the Einstein equations. This route leads us to the fact that, as equations for the fundamental variables internal to the

boundary hypersurface, the vanishing projections of the Einstein tensor are not identically satisfied by the solution of the initial-boundary value problem of the fundamental variables, but must be imposed as boundary conditions, and are consistent with the analysis of the initial-boundary value problem. As equations for the constraint variables, they represent consistent boundary conditions for the initial-boundary value problem of the constraints implied by the evolution equations. The latter is a consequence of the fact that the components of the Einstein tensor projected normally to the boundary are equivalent to the constraints when they are evaluated on solutions of the evolution equations.

The projection of the Einstein tensor normally to a timelike hypersurface thus gives rise to appropriate boundary conditions for the initial-boundary value problems of the fundamental variables and of the constraints *simultaneously.* Because such equations are internal to the boundary in the sense that they involve no second derivatives across the boundary, they require no information from outside, in exactly the same sense that the initial constraints require no information from before the initial slice. They function as a screen against the incoming flow of information, much of which being incompatible with the Einstein equations.

This contribution unifies results that have recently appeared separately in the literature [4, 5, 7, 6]. The presentation is meant to be accessible to a broader readership. The readers are referred to [4, 5, 7, 6] for details.

2 Preliminaries

Throughout we assume the metric of spacetime, g_{ab}, to be given in the $3+1$ notation with vanishing shift vector, that is:

$$ds^2 = g_{ab}dx^a dx^b = -\alpha^2 dt^2 + \gamma_{ij}dx^i dx^j \tag{1}$$

where $\alpha(t, x^k)$ is the lapse function and $\gamma_{ij}(t, x^k)$ is a three-dimensional Riemannian metric for the instantaneous time slices of spacetime. The assumption of vanishing shift vector is made for simplicity. The reader may assume that a non-vanishing shift vector will greatly increase the complexity of the calculations in the arguments presented here. The Einstein equations $G_{ab} = 0$ for the spacetime metric can be expressed in the ADM form [17]:

$$\dot{\gamma}_{ij} = -2\alpha K_{ij}, \tag{2}$$

$$\dot{K}_{ij} = \alpha\left(R_{ij} - 2K_{il}K^l{}_j + KK_{ij}\right) - D_i D_j \alpha\,, \tag{3}$$

with the constraints

$$\mathcal{C} \equiv -\frac{1}{2}\left(R - K_{ij}K^{ij} + K^2\right) = 0\,, \tag{4}$$

$$\mathcal{C}_i \equiv D_j K^j{}_i - D_i K = 0\,. \tag{5}$$

Here an overdot denotes a partial derivative with respect to the time coordinate ($\partial/\partial t$), indices are raised with the inverse metric γ^{ij}, D_i is the covariant three-derivative consistent with γ_{ij} , R_{ij} is the Ricci curvature tensor of γ_{ij}, R its Ricci scalar, K_{ij} is the extrinsic curvature of the slice at fixed value of t and $K \equiv \gamma^{ij} K_{ij}$. Expressed in terms of the Einstein tensor, the constraints (4)–(5) are related to specific components in the coordinates (t, x^i):

$$\mathcal{C} = -\alpha^2 G^{tt} \,, \tag{6}$$

$$\mathcal{C}_i = -\alpha \gamma_{ij} G^{jt} \,, \tag{7}$$

where (7) holds only for vanishing shift vector. The constraint character (the absence of second derivatives with respect to time) is a consequence of the fact that the only components of the Einstein tensor that appear in (6)–(7) have a contravariant index of value t. This follows from the Bianchi identities, $\nabla_a G^{ab} = 0$, via the following argument [16]. Writing the Bianchi identities explicitly in terms of the coordinates (t, x^i) we have

$$\partial_t G^{tb} = -\partial_x G^{xb} - \partial_y G^{yb} - \partial_z G^{zb} - {}^4\Gamma^a{}_{ac} G^{cb} - {}^4\Gamma^b{}_{ac} G^{ac} \tag{8}$$

where ${}^4\Gamma^c{}_{ab}$ are the Christoffel symbols of the spacetime metric g_{ab}. Manifestly, the right-hand side contains no time derivatives of any components of the spacetime metric of order higher than second. Therefore $\partial_t G^{tb}$ does not involve time derivatives of order higher than second, thus G^{tb} does not involve time derivatives of order higher than *first.* Linear combinations of the four components G^{tb} will also lack second time-derivatives of any components of the spacetime metric. Then G^{tb} or any four linearly independent combinations of them are, effectively, *constraints with respect to the time coordinate,* since the remaining six components of the Einstein tensor – represented in "unfolded" form in terms of γ_{ij} and K_{ij} by the twelve equations in the set (2)–(3) – all contain second time-derivatives of γ_{ij}. In geometric terms, denoting by n^b the unit normal to the slices of fixed value of t, given by $n^a = g^{ab} n_b = -\alpha g^{at} = \delta^a_t/\alpha$, one sees that $G_{ab} n^b = -\alpha G^t_a$, thus (6)–(7) are vanishing linear combinations of the projection of the Einstein tensor along the normal to the time slices, $G_{ab} n^b = 0$.

By a similar argument, the Bianchi identities also imply that vanishing linear combinations of the components of the Einstein tensor with a contravariant x index, that is $G^{ax} = 0$ are, effectively, constraints with respect to the x coordinate (likewise, $G^{ay} = 0$ and $G^{az} = 0$ are constraints with respect to the y and z coordinates, respectively). The components $G^{ax} = 0$ can be given a geometric interpretation in terms of the unit spacelike vector e^b normal to a constant$-x$ hypersurface, which is given by $e^a = g^{ab} e_b = g^{ax}/\sqrt{g^{xx}} = (0, \gamma^{ix}/\sqrt{\gamma^{xx}})$. We have that $G^{ax} = g^{ac} G^x_c = g^{ac} g^{bx} G_{cb} = g^{ac} \sqrt{g^{xx}} G_{cb} e^b$. Thus it is vanishing linear combinations of the components of the projection of the Einstein tensor along the normal to a constant$-x$ hypersurface that are the four *constraints with respect to the x coordinate,* that is: $G_{ab} e^b = 0$. Explicitly we have:

$$G^x_t = -\frac{1}{2}\gamma^{ix}\left((\ln\gamma)_{,it} - \gamma^{kl}\dot{\gamma}_{ik,l}\right) - KD^x\alpha + K^x_k D^k\alpha$$
$$+\alpha\left(\gamma^{kl}\Gamma^j{}_{kl}K^x_j + \gamma^{ix}\Gamma^j{}_{ik}K^k_j\right) , \tag{9}$$

$$G^x_x = \frac{\dot{K} - \dot{K}^x_x}{\alpha} - \frac{1}{2}(R + K^{ij}K_{ij} + K^2) + KK^x_x$$
$$+R^x_x + \frac{1}{\alpha}\left(D^jD_j\alpha - D^xD_x\alpha\right) , \tag{10}$$

$$G^x_y = -\frac{\dot{K}^x_y}{\alpha} + KK^x_y + R^x_y - \frac{1}{\alpha}D^xD_y\alpha \tag{11}$$

$$G^x_z = -\frac{\dot{K}^x_z}{\alpha} + KK^x_z + R^x_z - \frac{1}{\alpha}D^xD_z\alpha\ . \tag{12}$$

Here $\Gamma^k{}_{ij} = (1/2)\gamma^{kl}(\gamma_{il,j} + \gamma_{jl,i} - \gamma_{ij,l})$, and the time derivative of the components of the extrinsic curvature is applied after raising an index, that is: $\dot{K}^i_j \equiv (\gamma^{ik}K_{kj})_{,t}$. One sees that $G_{ab}e^b = 0$ are evolution equations for some of the fundamental variables. The difference with (2)–(3) is that $G_{ab}e^b = 0$ are internal to surfaces of fixed value of x.

With respect to the individual roles of all three sets of equations (2)–(3), (4)–(5) and (9)–(12), one may tentatively think that (4)–(5) must be solved for the values of γ_{ij} and K_{ij} at $t = 0$, and then γ_{ij} and K_{ij} at $t = dt$ would be obtained by (2)–(3) at any point x^i except at the boundary, where only four of the 12 fundamental variables can be obtained by the use of (9)–(12). But how to make sense of (2)–(3), (4)–(5) and (9)–(12) in terms of an initial-boundary value problem for γ_{ij} and K_{ij} is not straightforward, and we refrain from the task. In the next section, we use a first-order reduction of the ADM equations in order to find the role that (9)–(12) play in the framework of a first-order initial-boundary value problem for the Einstein equations.

3 The Components of the Projection $G_{ab}e^b = 0$ as Boundary Conditions

In this section we develop the notion that the projections $G_{ab}e^b = 0$ restrict the values of some fundamental variables that are usually thought to be arbitrary along the boundary for the purpose of generating a solution to the Einstein equations from constrained initial data. The normal projections to the boundary surface, thus, assume the role of choosers of boundary values, in analogy with the role of the projections normal to the initial slice as choosers of initial data. Intuitively, the boundary equations $G_{ab}e^b = 0$ select information that comes into the region of interest by flowing in through the boundaries. This concept is made precise in the following by resorting to the mathematical technology of strongly hyperbolic systems of partial differential equations [9].

3.1 Strongly Hyperbolic Formulations of the Einstein Equations

In order to have a system of equations that yields solutions to the Einstein equations and is well posed at the same time, a common strategy is to start by rewriting the ADM equations into first-order form by defining a set of 18 new variables, f_{kij}, symmetric in the pair (i, j), that are linearly independent combinations of the space-derivatives of the metric $\gamma_{ij,k}$. The new variables evolve according to equations derived from the ADM equation (2), so the ADM system is enlarged by 18 more equations. In that way, the ADM equations are automatically expanded into a 30-dimensional quasilinear system of partial differential equations of the first order,

$$\dot{\boldsymbol{u}} = \boldsymbol{A^i} \boldsymbol{u}_{,i} + \boldsymbol{b} \,, \tag{13}$$

where $\boldsymbol{u} = (\gamma_{ij}, K_{ij}, f_{kij})$ represents the set of 30 fundamental variables, $\boldsymbol{A^i} = \boldsymbol{A^i}(\boldsymbol{u}, t, x^i)$ are three 30-dimensional matrices and $\boldsymbol{b} = \boldsymbol{b}(\boldsymbol{u}, t, x^i)$ represents undifferentiated terms. The new first-order system of equations is equivalent to the Einstein equations if the map between $\gamma_{ij,k}$ and f_{kij} is written in the form

$$\mathcal{C}_{kij}(f_{kij}, \gamma_{ij,k}) = 0 \tag{14}$$

and included as 18 new constraints in addition to $\mathcal{C}$ and $\mathcal{C}_i$.

Such a first-order version of the Einstein equations is said to be *strongly hyperbolic* if and only if, for every arbitrary fixed covector ξ_i such that $\gamma^{ij}\xi_i\xi_j = 1$, the matrix $\boldsymbol{A} \equiv \boldsymbol{A^i}\xi_i$ has real eigenvalues and a complete set of eigenvectors [9]. It may be worth pointing out that it seems to be impossible to find a transformation between $\gamma_{ij,k}$ and f_{kij} that brings about strong hyperbolicity to the first-order reduction of the Einstein equations unless the evolution equations for f_{kij} that are derived from the ADM equation (2) are modified by the addition of terms proportional to the constraints [11, 12, 8].

For every spatial direction ξ_i, the existence of a complete set of eigenvectors of the principal symbol is equivalent to the existence of a (local) basis of traveling waves for the fundamental variables, referred to as the *characteristic fields*. For each ξ_i, the set of characteristic fields splits into three subsets: those that travel with zero speed, those that travel in the direction of ξ_i with some associated characteristic speed and those that travel in the direction opposite to ξ_i with their own characteristic speed. If ξ_i is the outward-pointing normal to a boundary, then the characteristic fields that travel in the direction of ξ_i are referred to as *outgoing fields,* and those that travel opposite to ξ_i are referred to as *incoming fields.* In what follows, the term "static fields" is used to denote the characteristic fields that travel with zero speed.

Because the characteristic fields are essentially traveling waves, their field values at some initial time propagate along the direction of travel at their associated characteristic speed. Therefore, the values of the static and outgoing characteristic fields at a timelike boundary are propagated from prescribed initial values in the interior and cannot be assigned arbitrarily. On the other

hand, the values of the incoming characteristic fields at the boundary "would come from outside" and must be prescribed in order for the problem to have a unique solution [9].

Summarizing, the initial-boundary-value problem of a generic strongly hyperbolic reduction of the Einstein equations requires the prescription of initial data satisfying $4 + 18 = 22$ constraints, and as many boundary values as incoming characteristic fields. Because the Einstein equations constitute a constrained system, at this point it is not completely clear whether the boundary values can be specified arbitrarily (much unlike the case of a generic unconstrained problem where the incoming boundary values are completely free). In fact, if it turns out that any of the four projections $G_{ab}e^b = 0$ given by (9)–(12) involve the incoming characteristic fields, then clearly those incoming characteristic fields involved *may not be assigned arbitrary boundary values,* but must be prescribed in accordance with (9)–(12). On the contrary, if no incoming characteristic fields appear in (9)–(12), then the required boundary values may be prescribed arbitrarily. In order to gain insight into the arbitrariness of the boundary values, in the next Subsection we specialize the discussion to a particular well-known reduction of the Einstein equations.

3.2 The Case of the Einstein–Christoffel Formulation

The Einstein-Christoffel (EC) formulation, due to Anderson and York [1], is a symmetric-hyperbolic first-order reduction of the Einstein equations that is achieved by taking three steps away from the ADM equations: the addition of a certain set of 18 first-order variables representing the space-derivatives of the metric, the addition of terms proportional to the constraints in the evolution equations of such new first-order variables, and the prescription of a lapse function proportional to the square root of the determinant of the metric, that is, $\alpha \equiv Q\sqrt{\gamma}$ with Q assumed arbitrarily prescribed a priori. The new (first-order) variables are the following set of linear combinations of the derivatives of the metric:

$$f_{kij} \equiv \Gamma_{(ij)k} + \gamma_{ki}\gamma^{lm}\Gamma_{[lj]m} + \gamma_{kj}\gamma^{lm}\Gamma_{[li]m} \,, \tag{15}$$

where ${\Gamma^k}_{ij}$ are the Christoffel symbols of γ_{ij}. With this definition and the choice of densitized lapse, the right-hand side of the ADM equation for the evolution of the extrinsic curvature, (3), reduces to a divergence with additional undifferentiated terms. Evolution equations for f_{kij} can be obtained by taking a time derivative of (15) and using the ADM equation for the evolution of the metric, (2), to eliminate the time derivatives in favor of space-derivatives of the extrinsic curvature in the right-hand side. By adding terms proportional to the vector constraint $\mathcal{C}_i$, the right-hand side of the evolution equation for f_{kij} reduces to a gradient of the extrinsic curvature. The pair of evolution equations for K_{ij} and f_{kij} represents thus a first-order reduction of wave equations with additional undifferentiated terms [10]:

$$\dot{K}_{ij} = -\alpha\gamma^{kl}\partial_l f_{kij} + \cdots \tag{16}$$

$$\dot{f}_{kij} = -\alpha\partial_k K_{ij} + \cdots \tag{17}$$

The constraints are

$$\begin{aligned}\mathcal{C} \equiv & -\frac{1}{2}\gamma^{ij}\gamma^{kl}\{2(\partial_k f_{ijl} - \partial_i f_{jkl}) + K_{ik}K_{jl} - K_{ij}K_{kl} \\ & +\gamma^{mn}[f_{ikm}(5f_{jln} - 6f_{ljn}) + 13f_{ikl}f_{jmn} \\ & +f_{ijk}(8f_{mln} - 20f_{lmn}]\} = 0 \end{aligned} \tag{18}$$

$$\begin{aligned}\mathcal{C}_i \equiv & -\gamma^{kl}\{\gamma^{mn}[K_{ik}(3f_{lmn} - 2f_{mnl}) - K_{km}f_{iln}] \\ & +\partial_i K_{kl} - \partial_k K_{il}\} = 0 \end{aligned} \tag{19}$$

$$\mathcal{C}_{kij} \equiv 2f_{kij} - 4\gamma^{lm}(f_{lm(i}\gamma_{j)k} - \gamma_{k(i}f_{j)lm}) - \partial_k\gamma_{ij} = 0 \tag{20}$$

and are to be imposed only on the initial data. Here $\mathcal{C}_{ijk}$ represent the definition of the additional 18 first-order variables f_{kij} and are obtained by inverting (15) for $\gamma_{ij,k}$.

The boundary equations, (9)–(12), take the following form when translated in terms of the EC variables [7]:

$$G^x_t = \dot{f}^x{}_k{}^k - \dot{f}^k{}_k{}^x + \cdots \tag{21}$$

$$\begin{aligned}G^x_x = & \frac{\dot{K} - \dot{K}^x_x}{\alpha} + \partial_z(f^z{}_z{}^z + 2f^z{}_x{}^x - f^{xz}{}_x + 3f^z{}_y{}^y - 2f^y{}_y{}^z) \\ & +\partial_y(f^y{}_y{}^y + 2f^y{}_x{}^x - f^{xy}{}_x + 3f^y{}_z{}^z - 2f^z{}_z{}^y) + \cdots \end{aligned} \tag{22}$$

$$\begin{aligned}G^x_y = & -\frac{\dot{K}^x_y}{\alpha} + \partial_z\left(f^x{}_y{}^z - f^z{}_y{}^x\right) \\ & +\partial_y\left(f^x{}_y{}^y - f^y{}_y{}^x - 3f^x{}_k{}^k + 2f^k{}_k{}^x\right) + \cdots \end{aligned} \tag{23}$$

$$\begin{aligned}G^x_z = & -\frac{\dot{K}^x_z}{\alpha} + \partial_y\left(f^x{}_z{}^y - f^y{}_z{}^x\right) \\ & +\partial_z\left(f^x{}_z{}^z - f^z{}_z{}^x - 3f^x{}_k{}^k + 2f^k{}_k{}^x\right) + \cdots \end{aligned} \tag{24}$$

where ... represent undifferentiated terms.

The EC equations (16)–(17) are symmetric hyperbolic (and thus strongly hyperbolic as well [9]). With respect to the unit vector $\xi^i \equiv \gamma^{xi}/\sqrt{\gamma^{xx}}$ which is normal to the boundary $x = x_0$ for the region $x \leq x_0$ there are 18 "static" characteristic fields (the six γ_{ij}, the six $f^y{}_i{}^j$ and the six $f^z{}_i{}^j$) and 12 characteristic fields traveling at the speed of light, of which six are incoming:

$${}^-U^j_i \equiv K^j_i - \frac{f^x{}_i{}^j}{\sqrt{\gamma^{xx}}} \tag{25}$$

and six are outgoing:

$${}^+U^j_i \equiv K^j_i + \frac{f^x{}_i{}^j}{\sqrt{\gamma^{xx}}}\ . \tag{26}$$

All outgoing and static characteristic fields are assumed to be determined by their initial values. The incoming fields, though, may be restricted by the boundary equations. It is our next task to figure out which incoming fields are constrained by the boundary equations, if any. Inverting for K_i^j and $f^x{}_i{}^j$ in terms of ${}^{\pm}U_i^j$ one has

$$K_i^j = \frac{1}{2}({}^{+}U_i^j + {}^{-}U_i^j)\,, \tag{27}$$

$$f^x{}_i{}^j = \frac{\sqrt{\gamma^{xx}}}{2}({}^{+}U_i^j - {}^{-}U_i^j)\,. \tag{28}$$

One can thus see that (23) and (24) prescribe the time derivatives of ${}^{-}U_y^x$ and ${}^{-}U_z^x$ in terms of the time derivatives of ${}^{+}U_y^x$ and ${}^{+}U_z^x$, respectively. This means that ${}^{-}U_y^x$ and ${}^{-}U_z^x$ cannot be assigned arbitrary boundary values in order to generate a solution of the Einstein equations. Their values can be calculated from knowledge of the outgoing fields ${}^{+}U_y^x$ and ${}^{+}U_z^x$ and previous knowledge of other fields, by using (23) and (24) as evolution equations along the boundary.

One can further see that (22) and (21) both involve the time derivative of the same combinations of incoming characteristic fields, explicitly: ${}^{-}U_y^y + {}^{-}U_z^z$. In the first place, this means that ${}^{-}U_y^y + {}^{-}U_z^z$ cannot be assigned arbitrary values along the boundary. Secondly, both equations also involve the time derivative of the same combination of outgoing fields, namely: ${}^{+}U_y^y + {}^{+}U_z^z$. Thus both equations taken together up to linear combinations are equivalent to: **(1)** a boundary prescription for ${}^{-}U_y^y + {}^{-}U_z^z$, and **(2)** a consistency condition for ${}^{+}U_y^y + {}^{+}U_z^z$. The boundary prescription for ${}^{-}U_y^y + {}^{-}U_z^z$ can be any linear combination of (22) and (21) except $\alpha G_x^x + G_t^x/\sqrt{\gamma^{xx}} = 0$, in principle.

The combination $\alpha G_x^x + G_t^x/\sqrt{\gamma^{xx}} = 0$ can be taken as the consistency condition for ${}^{+}U_y^y + {}^{+}U_z^z$. That this is a consistency condition means that if the values of ${}^{+}U_y^y + {}^{+}U_z^z$ were calculated by using it as an evolution equation along the boundary, they should be identical to the values that would arise by characteristic propagation from the initial slice. It also means that, in principle, $\alpha G_x^x + G_t^x/\sqrt{\gamma^{xx}} = 0$ can (but does not have to) be ignored as a boundary equation. The question of whether or not to ignore $\alpha G_x^x + G_t^x/\sqrt{\gamma^{xx}} = 0$ as a boundary equation remains an open problem at this time.

In all, thus, three out of the four components $G_a^x = 0$ prescribe boundary values for three out of the six incoming fields. That leaves three incoming fields for which the boundary values can be assigned arbitrarily without being inconsistent with the Einstein equations. Up to linear combinations, they are ${}^{-}U_x^x$, ${}^{-}U_z^y$ and ${}^{-}U_y^y - {}^{-}U_z^z$. These are in one-to-one correspondence with three components of the extrinsic curvature, namely K_x^x, K_z^y and $K_y^y - K_z^z$. Of these, the last two can be identified with the transverse traceless part of the extrinsic curvature.

One of the consequences of this analysis is that some widely used choices of boundary conditions in numerical relativity are inconsistent with the Einstein

equations along the boundary. For instance, the *freezing* boundary conditions, where the time derivatives of *all* incoming fields are set to zero (${}^{-}\dot{U}^i_j = 0$), are clearly inconsistent with all four of (21)–(24), because of the presence of all the undifferentiated terms. However, one could implement "Einstein boundary conditions" for ${}^{-}U^x_y$, ${}^{-}U^x_z$ and ${}^{-}U^y_y + {}^{-}U^z_z$ by imposing, for instance, (23), (24) and (21), and freezing boundary conditions on the remaining three incoming fields. As another instance, notice that (21)–(24) are also inconsistent with imposing outgoing radiation conditions of the form ${}^{-}U^i_j = {}^{+}U^i_j$ on all incoming fields. There may be instances of outgoing radiation conditions that are different from ${}^{-}U^i_j = {}^{+}U^i_j$ because the term "outgoing radiation conditions" is used somewhat loosely accross the field. In such cases, the form of the "outgoing radiation conditions" must be contrasted to (21)–(24) in order to find out whether they are consistent with the Einstein equations along the boundary.

4 The Projection $G_{ab}e^b = 0$ in Relation to the Propagation of the Constraints

The discussion so far makes the point that the projection of the Einstein tensor normally to a timelike boundary gives rise to a set of consistent boundary conditions for the initial-boundary value problem of the Einstein equations. Indeed, we have demonstrated that the assignment of arbitrary boundary values to the incoming fields directly results in the failure to satisfy the Einstein equations along the boundary surface. In the following, we establish the greater consequences of imposing $G_{ab}e^b = 0$ on the boundary. We will make the point that, essentially, the use of $G_{ab}e^b = 0$ as boundary conditions for the fundamental variables is instrumental for the purposes of guaranteeing that the solution of the evolution equations satisfies the constraints at any time.

4.1 The Case of the ADM Equations

Even though the ADM equations themselves are not strongly hyperbolic in the standard sense – discussed in Subsect. 3.1 -, they imply a strongly hyperbolic system of evolution equations for the constraint quantities $\mathcal{C}, \mathcal{C}_i$ as functions of the point. These evolution equations for the constraints are obtained by taking a time-derivative of $\mathcal{C}(\gamma_{ij}, K_{ij})$ and $\mathcal{C}_i(\gamma_{ij}, K_{ij})$, and using the evolution equations (2)–(3) in order to eliminate $\dot{\gamma}_{ij}$ and $\dot{K}_{ij}$ in favor of space derivatives of γ_{ij} and K_{ij}. The substitution allows for the highest derivative terms in the resulting equations to be expressible in terms of first derivatives of the constraint quantities themselves, resulting in a closed system for $\mathcal{C}, \mathcal{C}_i$ having the form:

$$\dot{\mathcal{C}} = \alpha \partial^i \mathcal{C}_i + \cdots \tag{29}$$

$$\dot{\mathcal{C}}_i = \alpha \partial_i \mathcal{C} + \cdots \tag{30}$$

where ... denote undifferentiated terms. This system of equations has two characteristic speeds: 0 and $\pm\alpha$. With respect to the unit vector $\xi^i = \gamma^{ix}/\sqrt{\gamma^{xx}}$, normal to the boundary, the characteristic fields that travel with zero speed are $\mathcal{C}^y$ and $\mathcal{C}^z$. The characteristic fields that travel with non-zero speed $\pm\alpha$ are, respectively, ${}^{\pm}\mathcal{C} \equiv \mathcal{C} \pm \mathcal{C}^x/\sqrt{\gamma^{xx}}$. The characteristic constraint ${}^{+}\mathcal{C}$ is outgoing at the boundary, whereas ${}^{-}\mathcal{C}$ is incoming at the boundary. Because there is one incoming characteristic constraint, the initial-boundary value problem of (29)–(30) requires one boundary prescription in order for a unique solution to exist for every set of initial data. In other words, even if the constraints are set to zero initially, they will not remain vanishing at all times unless a "constraint-preserving" condition is imposed on the boundary. It is at this crucial point that the initial value problem and the initial-boundary value problem of the Einstein equations differ with respect to the propagation of the constraints. In the case of the initial value problem, constrained initial data uniquely pick out solutions of the evolution equations that satisfy the constraints at all times. In the case of the initial-boundary value problem, constrained initial data alone are not sufficient to pick out the solutions that satisfy the constraints at all time, but constraint-preserving boundary data must be specified as well.

The initial-boundary value problem of (29)–(30) is much more general than one needs for the Einstein equations. In fact, (29)–(30) admit any initial values, whereas the Einstein equations require *vanishing* initial values for the constraint quantities. So, one is really interested in the initial-boundary value problem of (29)–(30) with *homogeneous* initial values and boundary conditions. The boundary prescription does not have to consist of setting ${}^{-}\mathcal{C} = 0$, but could be setting ${}^{-}\mathcal{C}$ as any linear combination of ${}^{+}\mathcal{C}, \mathcal{C}^y$ and $\mathcal{C}^z$: as the "static"and outgoing constraints are set to zero initially and propagate towards the boundary, the linear combination is equivalent to prescribing a vanishing value for the incoming constraint.

Our proposed boundary conditions for the fundamental variables are (9)–(12), so our immediate interest is to figure out how they relate to the constraint quantities $\mathcal{C}$ and $\mathcal{C}_i$. The relationship may not be expected to be direct, because time derivatives of the fundamental variables occur in G_a^x, but do not occur in $\mathcal{C}$ or $\mathcal{C}_i$. But since we will eventually be interested only in the values of the fundamental variables as they satisfy the evolution equations, we may use the evolution equations to eliminate the time derivatives that occur in G_a^x and figure out the relationship of whatever is left to the constraints. The relationship will thus be one of *equivalence modulo the evolution equations.*

To start with, by simple inspection one can see that if the evolution equations for the extrinsic curvature are used to eliminate $\dot{K}_y^x$ and $\dot{K}_z^x$ from G_y^x and G_z^x, the corresponding boundary equations reduce to naught. That is: the boundary equations $G_y^x = 0$ and $G_z^x = 0$ are two of the six evolution

equations and are not related to the constraints. But if one uses the evolution equations for the extrinsic curvature in order to eliminate $\dot{K}$ and $\dot{K}^x_x$ from G^x_x, one is left with a series of terms that is exactly identical to the scalar constraint $\mathcal{C}$ up to an overall minus sign. One may interpret this fact as indicating that $G^x_x = 0$ is *equivalent* to imposing the vanishing of the scalar constraint on the boundary. Finally, using the evolution equation for the metric to eliminate $\dot{\gamma}_{ij}$ from G^x_t one is left with a series of terms that is identical to the combination $\mathcal{C}^x \equiv \gamma^{xi}\mathcal{C}_i$ up to an overall factor of α. This can be interpreted as indicating that $G^x_t = 0$ is equivalent to imposing $\mathcal{C}^x = 0$ on the boundary. In summary:

$$G^x_t \sim \alpha \mathcal{C}^x \tag{31}$$

$$G^x_y \sim 0 \tag{32}$$

$$G^x_z \sim 0 \tag{33}$$

$$G^x_x \sim -\mathcal{C}\ . \tag{34}$$

One can thus see that by using either $G^x_t = 0$ or $G^x_x = 0$ as a boundary condition for the fundamental variables, one is effectively imposing a consistent boundary condition on the problem of propagation of the constraints, of the form ${}^-\mathcal{C} = {}^+\mathcal{C}$ or ${}^-\mathcal{C} = -{}^+\mathcal{C}$, respectively. In fact, one could use any linear combination of $G^x_t = 0$ and $G^x_x = 0$ *except* $G^x_t - (\alpha/\sqrt{\gamma^{xx}})G^x_x = 0$ as the boundary prescription necessary to enforce the propagation of the constraints. On the other hand, $G^x_t - (\alpha/\sqrt{\gamma^{xx}})G^x_x = 0$ in itself is equivalent to ${}^+\mathcal{C} = 0$ on the boundary, which should be identically satisfied if all the constraints are set to zero on the initial slice. Thus $G^x_t - (\alpha/\sqrt{\gamma^{xx}})G^x_x = 0$ is not really a boundary prescription but a consistency condition on the boundary.

4.2 The Case of the Einstein–Christoffel Formulation

In the case of the EC formulation, we have already shown that three of the four equations $G^x_a = 0$ impose non-trivial boundary conditions on three of the six incoming fields, whereas the fourth one represents a consistency condition on an outgoing field. Here we develop the relationship of $G^x_a = 0$ to the constraints in the EC framework. The aim is to show that the three non-trivial boundary equations of the set $G^x_a = 0$ are necessary and sufficient for the purpose of picking a solution of the EC equations that satisfies the constraints at all times.

We start by looking at the auxiliary system of propagation of the constraints. By taking a time derivative of the constraints (18)–(20) and using the evolution equations (16)–(17) in order to eliminate time derivatives of the fundamental variables we have explicitly:

$$\dot{\mathcal{C}} = \alpha \partial^i \mathcal{C}_i + \cdots \tag{35}$$

$$\dot{\mathcal{C}}_i = \frac{1}{2}\alpha \partial^k \left(\partial_i \mathcal{C}_{kl}{}^l - \partial_l \mathcal{C}_{ki}{}^l + \partial_k \mathcal{C}_{li}{}^l - \partial_k \mathcal{C}_{il}{}^l \right) - \alpha \partial_i \mathcal{C} + \cdots \tag{36}$$

$$\dot{\mathcal{C}}_{kij} = \ldots \tag{37}$$

where ... denote undifferentiated terms. The system of auxiliary evolution equations (35)–(37) for the 22 constraints has second-derivatives in the right-hand side. However, it can be reduced to first differential order in the usual manner, by adding new "constraint" variables that are space-derivatives of the constraints, such as

$$\mathcal{C}_{lkij} \equiv \frac{1}{2}\left(\partial_l \mathcal{C}_{kij} - \partial_k \mathcal{C}_{lij}\right) \ . \tag{38}$$

The introduction of such new constraint quantities enlarges the system (35)–(37) to 40 variables in all, and casts it into the following form:

$$\dot{\mathcal{C}} = \alpha \partial^i \mathcal{C}_i + \cdots \tag{39}$$

$$\dot{\mathcal{C}}_i = -\alpha\left(\partial_i \mathcal{C} + \partial^k \mathcal{C}_{lki}{}^l + \partial^k \mathcal{C}_{kil}{}^l\right) + \cdots \tag{40}$$

$$\dot{\mathcal{C}}_{kij} = \ldots \tag{41}$$

$$\dot{\mathcal{C}}_{lkij} = \alpha\left(\gamma_{ki}\partial_l \mathcal{C}_j + \gamma_{kj}\partial_l \mathcal{C}_i - \gamma_{li}\partial_k \mathcal{C}_j - \gamma_{lj}\partial_k \mathcal{C}_i\right) + \cdots \tag{42}$$

It is useful to point out that, if one chooses to do so, the new "constraint" variables $\mathcal{C}_{lkij}$ can be expressed in terms of the fundamental variables of the EC system, in which case they read:

$$\begin{aligned} \mathcal{C}_{lkij} = {} & \partial_l(f_{kij} - 2\gamma^{nm}(f_{nm(i}\gamma_{j)k} - \gamma_{k(i}f_{j)nm})) \\ & -\partial_k(f_{lij} - 2\gamma^{nm}(f_{nm(i}\gamma_{j)l} - \gamma_{l(i}f_{j)nm})) \ , \end{aligned} \tag{43}$$

and turn out to be identical to the "integrability conditions" $0 = 1/2(\partial_l \gamma_{ij,k} - \partial_k \gamma_{ij,l})$.

The immediate benefit of introducing the first-order constraint variables is that the system (39)–(42) is well posed in the sense that it is strongly hyperbolic with characteristic speeds of 0 (multiplicity 34), $+\alpha$ (multiplicity 3) and $-\alpha$ (multiplicity 3). This means that, with respect to the (outer) boundary at $x = x_0$, three characteristic constraint variables are incoming, three are outgoing and all the others are "static". The six non-static characteristic constraints, which we denote by ${}^{\pm}\mathcal{Z}_i$ (with + for outgoing and − for incoming), are explicitly:

$${}^{\pm}\mathcal{Z}_x = \mathcal{C}_x \pm \frac{1}{\sqrt{\gamma^{xx}}}\left(\mathcal{C} + \mathcal{C}^{kx}{}_{xk} + \mathcal{C}^{x}{}_{xk}{}^{k}\right) \ , \tag{44}$$

$${}^{\pm}\mathcal{Z}_y = \mathcal{C}_y \pm \frac{1}{\sqrt{\gamma^{xx}}}\left(\mathcal{C}^{kx}{}_{yk} + \mathcal{C}^{x}{}_{yk}{}^{k}\right) \ , \tag{45}$$

$${}^{\pm}\mathcal{Z}_z = \mathcal{C}_z \pm \frac{1}{\sqrt{\gamma^{xx}}}\left(\mathcal{C}^{kx}{}_{zk} + \mathcal{C}^{x}{}_{zk}{}^{k}\right) \ . \tag{46}$$

As in the ADM case, we are only interested in *homogeneous* initial and boundary conditions for the propagation of the constraints. Because three characteristic constraints are incoming at the boundary, the system of equations for

the propagation of the constraints requires three (and only three) boundary conditions, which can, for instance, take the form of linear combinations such as ${}^-\mathcal{Z}_i = L_i^{j\,+}\mathcal{Z}_j$ with coefficients L_i^j.

So the number of boundary conditions required by the propagation of the constraints (39)–(42) is the same as the number of non-trivial boundary equations $G_b^x = 0$ that impose restrictions on the boundary values of the fundamental variables of the EC equations (16)–(17). It is reasonable to suppose that the three incoming constraints are related to the three components of G_b^x that act as non-trivial boundary equations. Since G_b^x contain time derivatives but the cosntraints do not, the relationship will be one of equivalence modulo the evolution, as in the ADM case.

In principle, in order to find the relationship between G_b^x and the constraints one needs only to use the evolution equations (16)–(17) in order to eliminate the time derivatives that appear in (21)–(24), and then regroup the remainder into terms that are proportional to the constraints. The result is as follows [7]:

$$G_t^x \sim \alpha \mathcal{C}^x \,, \tag{47}$$

$$G_y^x \sim \mathcal{C}^{kx}{}_{yk} + \mathcal{C}^x{}_{yk}{}^k \,, \tag{48}$$

$$G_z^x \sim \mathcal{C}^{kx}{}_{zk} + \mathcal{C}^x{}_{zk}{}^k \,, \tag{49}$$

$$G_x^x \sim \mathcal{C} + \mathcal{C}^{kx}{}_{xk} + \mathcal{C}^x{}_{xk}{}^k \,. \tag{50}$$

One can see that the constraints involved in (47)–(50) are the same as those involved in (44)–(46), so we can express the right hand sides of (47)–(50) in terms of the non-static characteristic constraints:

$$G_t^x \sim \alpha \frac{\gamma^{xi}}{2} ({}^+\mathcal{Z}_i + {}^-\mathcal{Z}_i) \,, \tag{51}$$

$$G_y^x \sim \frac{\sqrt{\gamma^{xx}}}{2} ({}^+\mathcal{Z}_y - {}^-\mathcal{Z}_y) \,, \tag{52}$$

$$G_z^x \sim \frac{\sqrt{\gamma^{xx}}}{2} ({}^+\mathcal{Z}_z - {}^-\mathcal{Z}_z) \,, \tag{53}$$

$$G_x^x \sim \frac{\sqrt{\gamma^{xx}}}{2} ({}^+\mathcal{Z}_x - {}^-\mathcal{Z}_x) \,. \tag{54}$$

Of the four equations $G_b^x = 0$ one only needs three that are linearly independent combinations of the three incoming constraints. There is a number of ways to pick the three. One way is already suggested by the initial-boundary value problem of the EC equations in Subsect. 3.2, namely: to set $G_y^x = 0$, $G_z^x = 0$ and $aG_t^x + bG_x^x = 0$ for any values of a and b except $\alpha G_x^x + G_t^x/\sqrt{\gamma^{xx}} = 0$. This would be equivalent to setting

$${}^-\mathcal{Z}_i = L_i^{j\,+}\mathcal{Z}_j \,. \tag{55}$$

These are appropriate boundary conditions for the propagation of the constraints.

The fourth equation in the set, namely $\alpha G^x_x + G^x_t/\sqrt{\gamma^{xx}} = 0$, is equivalent to a linear relationship among purely outgoing constraints:

$$B^{j+}\mathcal{Z}_j = 0\,. \tag{56}$$

This equation is redundant to the problem of propagation of the constraints, being entirely consistent with the prescription of vanishing initial data.

The point is thus made that three out of the four components of the projection $G_{ab}e^b = 0$ are, in a sense, appropriate boundary conditions for the propagation of the constraints and, at the same time, that they are appropriate boundary conditions for the fundamental variables of the EC equations. The boundary conditions that arise from $G_{ab}e^b = 0$ can thus be interpreted as being constraint-preserving.

5 Concluding Remarks

Summarizing, the projections $G_{ab}e^b = 0$ have a significant role to play in the initial-boundary value problem of the Einstein equations, which is hereby illustrated in the case of the Einstein-Christoffel formulation. Just as the constraints $G_{ab}n^b = 0$ are necessary to weed out initial data that would lead to a four-dimensional metric that is inconsistent with the Einstein equations, the boundary equations $G_{ab}e^b = 0$ screen out boundary data that would flow in from outside in the form of the values of three incoming fields. By picking the values of the three incoming fields involved so that $G^x_t = G^x_y = G^x_z = 0$, for instance, one guarantees that the Einstein equations are satisfied at the boundary. Then, as the incoming fields travel towards the interior, they "carry", in a sense, the vanishing of G^x_t, G^x_y ad G^x_z to the region where the evolution equations are satisfied by construction. Since G^x_t, G^x_y and G^x_z differ from the incoming constraints only by terms that are proportional to the evolution equations, their vanishing is equivalent to the vanishing of the constraints wherever the evolution equations are satisfied. Thus, by setting $G^x_t = G^x_y = G^x_z = 0$ along the boundary, the constraints are enforced in the interior.

Consider the case that the boundary value problem arises from a Cauchy problem simply by choosing to restrict the problem to a bounded sector of the Cauchy surface. That is: given a problem where the initial data completely determine the solution, restrict attention to a bounded region of the initial slice. The solution everywhere, inside and outside of the artificial boundary created by the fiducial limits on the initial slice, is found entirely from global initial data satisfying the constraints. But, starting from the Cauchy surface, some of the constraints propagate into the artificial region from the values of initial data given outside of the region of interest. The values of the incoming fundamental variables along the artificial boundary must consequently be "constrained", since they arise from data that are initially constrained

in the part of the Cauchy surface that lies outside of the fiducial boundary. The realization of the constrained nature of the incoming fields takes place in the form of the boundary equations $G_{ab}e^b = 0$. We can thus think of $G_{ab}e^b = 0$ as a sort of time-like representation of the constraints, screening out of the artificially bounded region the same information that the constraints would screen out in the sector of the Cauchy surface outside of the fiducial boundary. By using $G_{ab}e^b = 0$ as boundary conditions in an artificially bounded problem, one is thus simply substituting some Cauchy data that one chooses to disregard with boundary data in an equivalent fashion. By virtue of their relationship with the incoming constraints, the boundary equations $G_{ab}e^b = 0$ realize along the boundary the selection mechanism that the constraints enforce on a Cauchy slice.

We conclude with a, perhaps far from complete, list of remarks and open questions.

In the first place, the most outstanding open question is whether there is a choice of linear combinations of the four components of $G_{ab}e^b = 0$ as boundary conditions that leads to a well-posed initial-boundary value problem in the sense that the solution will depend smoothly on the initial data. The question will remain wide open as long as the well-posedness of non-linear initial-boundary value problems remains a relatively unaddressed issue in the field of partial differential equations. Yet, in the linearization around Minkowski space, some results concerning well-posedness of the initial-boundary value problem with "Einstein boundary conditions" are available, as follows. It is known that the linearized EC evolution equations with constrained initial data and with $G^x_t = G^x_y = G^x_z = 0$ along the boundary can be trivially complemented with three more boundary conditions (in order to prescribe the remaining three incoming fields) in a way conducent to a well-posed initial-boundary value problem [13].

It is also worthwhile pointing out that we have hereby proved that $G^x_t = G^x_y = G^x_z = 0$ are identical to the "constraint preserving boundary conditions of the Neumann type" as defined in [3] but as applied to the fully nonlinear EC case (which are well-posed in the linearization). In fact, from this work it can be deduced that among all "constraint-preserving" boundary conditions – by which we mean all linear combinations of the form ${}^-\mathcal{Z}_i = L_i^{j+}\mathcal{Z}_j$ -, a subset are well posed (at least in the linearization), another subset are Einstein boundary conditions – by which we mean $G_{ab}e^b = 0$ up to linear combinations -, and the prescription $G^x_t = G^x_y = G^x_z = 0$ lies at the intersection which is strictly contained in both subsets. In terms of practicality, in all cases where both prescriptions of boundary data coincide, writing them out as projections of the Einstein tensor normal to the boundary surface greatly shortcuts the cumbersome procedure associated with the constraint-preserving scheme. An open question remains as to the meaning of the boundary conditions of the form ${}^-\mathcal{Z}_i = L_i^{j+}\mathcal{Z}_j$ that are not equivalent to projections of the Einstein

tensor (such as the "well-posed constraint-preserving boundary conditions of the Dirichlet type" of [3]).

It is important to emphasize that the significance of $G_{ab}e^b = 0$ to the initial-boundary value problem is independent of the formulation used. We expect that a similar analysis and conclusions can be carried out in the case of any strongly hyperbolic formulation of the Einstein equations. In fact, we have applied this analysis to the case of a strongly hyperbolic first-order reduction of the BSSN formulation [2] as well, obtaining very similar conclusions [6].

In closing, we speculate on the relevance of this work to numerical relativity. On the one hand, we have demonstrated that the prescription of boundary conditions is intimately connected with the propagation of the constraints in a fundamental way. This contradicts the – until recently – widespread practice in numerical relativity of giving arbitrary prescriptions of boundary data, shaped only by demands of numerical advantage. On the other hand, it has been observed [10] that the control of constraint violations is crucial for the purposes of extending the run time of numerical simulations of gravitational waves emitted by the collapse of binary black hole systems. The use of $G_{ab}e^b = 0$ as boundary conditions could thus play a role in improving both the precision and the accuracy of numerical simulations. It can be concluded that the use of $G_{ab}e^b = 0$ as boundary conditions has the potential for a significant and widespread impact in the future of the gravitational wave program.

Acknowledgements

This work was supported by the National Science Foundation under grants No. PHY-0244752 to Duquesne University and No. PHY-0135390 to Carnegie Mellon University. Additionally, Simonetta Frittelli is grateful to the Wilhelm und Else Heraeus-Stiftung and to Universität Tübingen for hospitality and support.

References

1. Arlen Anderson, James W. York: Fixing Einstein's equations. Phys. Rev. Lett. **82**, 4384 (1999) 211
2. Thomas W. Baumgarte, Stuart L. Shapiro: On the numerical integration of Einstein's field equations. Phys. Rev. D **59**, 024007 (1999) 221
3. Gioel Calabrese, Jorge Pullin, Oscar Reula, Olivier Sarbach, Manuel Tiglio: Well posed constraint preserving boundary conditions for the linearized Einstein equations. Commun. Math. Phys. **240**, 377 (2003) 220, 221
4. Simonetta Frittelli, Roberto Gómez: Boundary conditions for hyperbolic formulations of the Einstein equations. Class. Quantum Grav. **20**, 2379 (2003) 207

5. Simonetta Frittelli, Roberto Gómez: Einstein boundary conditions for the 3+1 Einstein equations. Phys. Rev. D **68**, 044014 (2003) 207
6. Simonetta Frittelli, Roberto Gómez: Einstein boundary conditions for the Einstein equations in the conformal-traceless decomposition. Phys. Rev. D **70**, 064008 (2004) 207, 221
7. Simonetta Frittelli, Roberto Gómez: Einstein boundary conditions in relation to constraint propagation for the initial-boundary value problem of the Einstein equations. Phys. Rev. D **69**, 124020 (2004) 207, 212, 218
8. Simonetta Frittelli: Potential for ill-posedness in several second-order formulations of the Einstein equations. Phys. Rev. D **70**, 044029 (2004) 210
9. Bertil Gustaffson, Heinz-Otto Kreiss, Joseph Oliger: *Time-Dependent Problems and Difference Methods* (Wiley, New York 1995) 209, 210, 211, 212
10. Lawrence E. Kidder, Mark A. Scheel, Saul A. Teukolsky, Eric D. Carlson, Gregory B. Cook: Black hole evolution by spectral methods. Phys. Rev. D **62**, 084032 (2000) 211, 221
11. Lawrence E. Kidder, Mark A. Scheel, Saul A. Teukolsky: Extending the lifetime of 3d black hole computations with a new hyperbolic system of evolution equations. Phys. Rev. D **64**, 064017 (2001) 210
12. Gabriel Nagy, Omar Ortiz, Oscar A. Reula: Strongly hyperbolic second order Einstein's evolution equations. Phys. Rev. D **70**, 044102 (2004) 210
13. Olivier Sarbach, Gioel Calabrese: Detecting ill-posed boundary conditions in general relativity. J. Math. Phys. **44**, 3888 (2003) 220
14. J. M. Stewart: The Cauchy problem and the initial boundary value problem in numerical relativity. Class. Quantum Grav. **15**, 2865 (1998) 206
15. Robert M. Wald: *General Relativity* (University of Chicago Press, Chicago 1984) 205
16. Steven Weinberg: *Gravitation and Cosmoloy. Principles and Applications of the General Theory of Relativity* (John Wiley & Sons, New York 1971) 205, 208
17. James W. York: Kinematics and dynamics of general relativity. In: *Sources of Gravitational Radiation*, ed by Larry Smarr (Cambridge University Press, Cambridge 1979) 205, 207

Recent Analytical and Numerical Techniques Applied to the Einstein Equations

Dave Neilsen[1,2], Luis Lehner[1], Olivier Sarbach[1,3], Manuel Tiglio[1,4,5]

[1] Department of Physics & Astronomy, Louisiana State University, Baton Rouge, LA 70803, USA
dneils1@lsu.edu
lehner@lsu.edu
sarbach@phys.lsu.edu
tiglio@phys.lsu.edu
[2] Department of Physics & Astronomy, Brigham Young University, Provo, UT 84602, USA
[3] Theoretical Astrophysics 130-33, California Institute of Technology, Pasadena, CA 91125, USA
[4] Center for Computation and Technology, Louisiana State University, Baton Rouge, LA 70803, USA
[5] Center for Radiophysics and Space Research, Cornell University, Ithaca, NY 14853

Abstract. Combining deeper insight of Einstein's equations with sophisticated numerical techniques promises the ability to construct accurate numerical implementations of these equations. We illustrate this in two examples, the numerical evolution of "bubble" and single black hole space-times. The former is chosen to demonstrate how accurate numerical solutions can answer open questions and even reveal unexpected phenomena. The latter illustrates some of the difficulties encountered in three-dimensional black hole simulations, and presents some possible remedies.

1 Introduction

Extracting the full physical content from Einstein's equations has proven to be a difficult task. The complexity of these equations has allowed researchers only a peek into the rich phenomenology of the theory by assuming special symmetries and reductions. Computational methods, however, are opening a new window into the theory. To realize the full utility of computational solutions in exploring Einstein's equations, several questions must first be addressed. Namely, a deeper understanding of the system of equations and its boundary conditions, the development and use of more refined numerical techniques and an efficient use of the available computational resources.

In recent years, considerable advances have been made in some of these issues, allowing for the analysis of complex physical systems which arguably must be tackled numerically. In the present article we highlight some recent

D. Neilsen et al.: *Recent Analytical and Numerical Techniques Applied to the Einstein Equations*, Lect. Notes Phys. **692**, 223–249 (2006)
www.springerlink.com

analytical and numerical techniques and apply them to two practical applications. The first application is the *numerical evolution of bubble space-times in five-dimensional Kaluza-Klein theory.* We study their dynamical behavior, the validity of cosmic censorship – in a set-up which a priori would appear promising to give rise to violations of the conjecture – and reveal the existence of critical phenomena. As a second application, we discuss the *numerical evolution of single black hole space-times.* Here we consider some analytical and numerical difficulties in modeling these systems accurately. We discuss a method to alleviate some of these problems, and present tests to demonstrate the promise of this method.

2 Analytical and Numerical Tools

In the Cauchy formulation of General Relativity, Einstein's field equations are split into evolution and constraint equations. Numerical solutions are found by specifying data on an initial space-like slice, subject to the constraints, and by integrating the evolution equations to obtain the future development of the data. Owing to finite computer resources, one is forced to use finite, and, in practice, rather small computational domains to discretize the problem. This raises several important issues.

The fundamental property for any useful numerical solution is that the solution must converge to the continuum solution in the limit of infinite resolution. A prerequisite for a well-behaved numerical solution is a well-posed continuum formulation of the initial-boundary value problem. In certain cases, the well-posed continuum problem can then be used to construct stable numerical discretizations for which one can a priori guarantee convergence. In particular, this can be achieved for linear, first-order, symmetric hyperbolic systems with maximally dissipative boundary conditions [1, 2, 3]. This is briefly discussed in Sect. 2.1, for a detailed description and an extension to numerical relativity see Refs. [4, 5, 6, 7].

The application of these ideas in general relativity is, naturally, more complicated. First, Einstein's equations are nonlinear and so it is much harder to a priori prove convergence. However, a discretization that guarantees stability for the linearized equations should already be useful for the nonlinear equations, especially for those systems with smooth solutions as expected for the Einstein equations when written appropriately. This is because in a small enough neighborhood of any given space-like slice, the numerical solution can be modeled as a small amplitude perturbation of the continuum solution.

The constraint equations in general relativity bring additional complications and greatly restrict the freedom in specifying boundary and initial data. This is illustrated and further discussed in Sect. 2.2. Secttion 2.3 discusses issues regarding the stability of the constraint manifold. The manifold is invariant with respect to the flow defined by the evolution system in the

continuum problem. Numerically, however, small errors in the solution arising from truncation or roundoff error may lead to large constraint violations if the constraint manifold is unstable. Section 2.3 discusses a method for suppressing such rapid constraint violations.

2.1 Guidelines for a Stable Numerical Implementation

A simple numerical algorithm, or "recipe," can be followed to solve first order, linear symmetric hyperbolic equations with variable coefficients and maximally dissipative boundary conditions, for which stability can be guaranteed. It is based on finite difference approximations with spatial difference operators that satisfy the *summation by parts* (SBP) property. This property is a discrete analogue of *integration by parts*, which is used in the derivation of energy estimates, a key ingredient for obtaining a well posed formulation of the continuum problem. SBP allows to obtain similar energy estimates for the discrete problem.

Employ Spatial Difference Operators that Satisfy SBP on the Computational Domain. For the sake of simplicity, consider a set of linear, first order symmetric hyperbolic equations in the one-dimensional domain $x \in (a,b)$ which is discretized with points $x_j = a + j\Delta x$, $j = 0 \ldots N$, where $\Delta x = (b-a)/N$. Now let us introduce the discrete scalar product,

$$(u,v) := \Delta x \sum_{i,j=0}^{N} \sigma_{ij} u_i v_j \tag{1}$$

for some positive definite matrix with elements σ_{ij}, which in the continuum limit $\Delta x \to 0$ approaches the L_2 scalar product $\langle u,v\rangle := \int_a^b uv\,dx$. At the continuum level, the derivative operator d/dx and the scalar product satisfy the rule of integration by parts, i.e. $\langle du/dx, v\rangle + \langle u, dv/dx\rangle = uv|_a^b$; in the discrete case this is translated into a finite difference operator D which satisfies $(Du,v) + (u,Dv) = uv|_a^b$ and approaches d/dx in the continuum limit. The simplest difference operator and scalar product satisfying SBP are

$$\begin{array}{lll} Du = (u_{i+1} - u_i)/\Delta x\,, & \sigma_{00} = \frac{1}{2} & \text{for } i = 0 \\ Du = (u_{i+1} - u_{i-1})/(2\Delta x)\,, & \sigma_{ii} = 1 & \text{for } i = 1 \ldots N-1 \\ Du = (u_i - u_{i-1})/\Delta x\,, & \sigma_{NN} = \frac{1}{2} & \text{for } i = N \end{array} \tag{2}$$

where the scalar product is diagonal: $\sigma_{ij} = 0$ for $i \neq j$. Higher order operators satisfying SBP have been constructed by Strand [2]. Additionally, when dealing with non-trivial domains containing inner boundaries, additional complexities must be addressed to attain SBP, see [4]. The finite operator D is then used for the discretization of the spatial derivatives in the evolution equations, thus obtaining a semi-discrete system.

Impose Boundary Conditions via Orthogonal Projections [3]. This ensures the consistent treatment of the boundaries, guaranteeing the correct handling of modes propagating towards, from and tangential to the boundaries. An energy estimate can be obtained for the semi-discrete system.

Implement an Appropriate Time Integration Algorithm. The resulting semi-discrete system constitutes a large system of ODE's which can be numerically solved by using a time integrator that satisfies an energy estimate [8, 9].

Consider Adding Explicit Dissipation. It is well known that finite difference approximations do not adequately represent the highest frequency modes on a given grid, corresponding to the shortest possible wavelengths that can be represented on the grid. If the smallest spacing between points is Δ, the shortest wavelength is $\lambda_{min} = \Delta$ with the corresponding frequency $k_{max} = 2\pi/\lambda_{min}$. These modes can, and often do, travel in the wrong direction. For this reason, it is sometimes useful to add explicit numerical dissipation to rid the simulation of these modes in a way that is consistent with the continuum equation at hand. If finer grids are used, the effect of this dissipation becomes smaller and acts only on increasingly higher frequencies. The dissipation operators are constructed such that discrete energy estimates, obtained using SBP, are not spoiled. Explicit expressions for such dissipation operators are presented in [4].

To summarize, beginning with a well-posed initial-boundary value problem, we mimic the derivation of continuum energy estimates for the discrete problem using (1) spatial derivative operators satisfying summation by parts, (2) orthogonal projections to represent boundary conditions and (3) choosing an appropriate time integrator.

2.2 Constraint-Preserving Boundary Conditions

As discussed above, a numerical implementation of any system of partial differential equations necessarily involves boundaries. Unless periodic boundary conditions can be imposed, as is often the case for the evolution on compact domains without boundaries, one deals with an initial-boundary value problem, and thus has to face the question of how to specify boundary conditions. In theories that give rise to constraints, like general relativity, such conditions must be chosen carefully to ensure that the constraints propagate.

As a very simple illustration, consider the 1d wave equation $u_{,tt} = u_{,xx}$ on the half line $x > 0$. Let us reduce it to first order form by introducing the variables $f \equiv u_{,x}$ and $g \equiv u_{,t} - bu_{,x}$, with b a negative constant:

$$u_{,t} = b\, u_{,x} + g, \tag{3}$$

$$g_{,t} = -b\, g_{,x} + (1 - b^2) f_{,x}\ , \tag{4}$$

$$f_{,t} = g_{,x} + b\, f_{,x}\ . \tag{5}$$

At the boundary $x = 0$, the system has two ingoing fields, given by u and $v_{in} \equiv g + b\,f - f$, and one outgoing field. However, the ingoing fields cannot be given independently, as we see next. The constraint $C \equiv f - u_{,x} = 0$ propagates as $C_{,t} = b\,C_{,x}$ and so C is an ingoing field with respect to $x = 0$. Therefore, we have to impose the boundary condition $C = 0$ which implies the condition $u_{,x} = f$ on the main variables. We can replace this with a condition that is intrinsic to the boundary by using the evolution equation (3) in order to eliminate the x-derivative and obtain

$$u_{,t} = b\,f + g\ . \tag{6}$$

This equation provides an evolution equation for determining u at the boundary, which guarantees that the constraint $C = 0$ is preserved throughout evolution. It can be complemented by the Sommerfeld condition $v_{in} = 0$.

This simple example gives just a glimpse of the different issues involved in prescribing constraint-preserving boundary conditions. The case of Einsteins's field equations is more complicated; we refer the interested reader to [10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22]. A major difficulty is the fact that, in general, constraint-preserving boundary conditions do not have the form of maximal dissipative boundary conditions, and for this reason it has proven to be difficult to find well posed initial-boundary value formulations of Einstein's equations that preserve the constraints.

2.3 Dealing with "Too Many" Formulations. Parameters via Constraint Monitoring

Formulations of the Einstein equations are often cast in symmetric hyperbolic form by adding constraints to the evolution equations multiplied by parameters or space-time functions. The symmetric hyperbolicity condition partially restricts these parameters. However, considerable freedom in the formulation exists in choosing these free parameters (see, for instance, [23]). Analytically, when data are on the constraint surface, all allowed values for these parameters are equally valid. Off the constraint surface, however, different values of these parameters may be regarded as representing "different" theories. It is no surprise then that numerical simulations are sensitive to the values chosen for these parameters, as numerical data rarely are on the constraint surface. Unfortunately, the parameters in current simulations are proving to be extremely sensitive. Relatively small variations in these parameters (within the allowed range for a symmetric hyperbolic formulation) produce run times in simulations that vary over several orders of magnitude, as measured by an asymptotic observer.

Furthermore, the parameters are not unique. Values convenient for one physical problem might be inappropriate in another. Recently, a method to dynamically choose these parameters – promoted to functions of time – was introduced that naturally adapts to the physical problem under study [24].

Basically, one exploits the freedom in choosing these functions to control the growth rate of an energy norm for constraint violations. Since this norm is exactly zero analytically, this provides a guide to choosing the parameters that will drive the solution to one that satisfies the constraints. This method provides a *practical* solution to this problem of choosing parameters, although it may not be the most elegant solution. Ideally, one would like to understand how the growth rate of the solution depends on the these parameter values in order to choose them appropriately. This would require sharp growth estimates, however, which are still unavailable. While further understanding is gained in this front, this practical remedy can be of much help in present simulations. We summarize here the essential ideas of this method.

Consider a system of hyperbolic equations with constraint terms, C_c, written schematically as

$$\dot{u}_a = \sum_b A^b(u,t,\mathbf{x})\partial_b u_a + B_a(u,t,\mathbf{x}) + \sum_c \mu_{ac} C_c(u,\partial_j u) \;, \tag{7}$$

where u_a, B_a and C_c are vector valued functions, and μ_{ac} is a matrix (generally not square) that is a function of the space-time (C_c represents a vector function of general constraint variables). The indices $\{a,b,c\}$ range over each element of the vector or matrix functions, while the indices $\{i,j,k\}$ label points on a discrete grid. We define an *energy* or *norm* of the discrete constraint variables as

$$\mathcal{N}(t) = \frac{1}{2n_x n_y n_z} \sum_c \sum_{ijk} C_c(t)^2 \;, \tag{8}$$

where n_x, n_y, n_z are the number of points in each direction. The grid indices $\{i,j,k\}$ are suppressed to simplify the notation. The time derivative of the norm can be calculated using (7)

$$\dot{\mathcal{N}} = \mathcal{I}^{hom} + \mathrm{Tr}(\mu \mathcal{I}^{\mu}) \;, \tag{9}$$

and therefore can be known in closed form provided the matrix valued sums

$$\begin{aligned} \mathcal{I}^{hom} = & \sum_{ijk} \sum_{a,b} \frac{C_a}{n_x n_y n_z} \left[\frac{\partial C_a}{\partial u_b} + \sum_k \frac{\partial C_a}{\partial D_k u_a} D_k \right] \\ & \times \left[\sum_c (A^c D_c u_b) + B_b \right] \end{aligned} \tag{10}$$

$$\begin{aligned} \mathcal{I}^{\mu}_{bc} = & \sum_{ijk} \sum_{a} \frac{C_a}{n_x n_y n_z} \\ & \times \left[\frac{\partial C_a}{\partial u_b} + \sum_k \frac{\partial C_a}{\partial D_k u_b} D_k \right] C_c \end{aligned} \tag{11}$$

are computed during evolution. Here D_i is the discrete derivative approximation to ∂_i. We then use the dependence of the energy growth on the free constraint-functions to achieve some desired behavior for the constraints, i.e., solving (9) for μ_{ac}. For example, if we choose[1]

$$\dot{\mathcal{N}} = -a\mathcal{N}, \qquad a > 0\,, \tag{12}$$

any violation of the constraints will decay exponentially

$$\mathcal{N}(t + \triangle t) = \mathcal{N}(t)e^{-a\triangle t}\,. \tag{13}$$

As discussed in [24], one good option among many others seems to be choosing a tolerance value, T, for the norm of the constraints that is close to the initial discrete value, and solving for μ_{ac} such that the constraints decay to this tolerance value after a given relaxation time. This can be done by adopting an a such that after some time $\tau \equiv n_a \triangle t$ the constraints have the value T. Replacing $\mathcal{N}(t + \triangle t)$ by T in equation (13) and solving for a gives

$$a(t) = -\frac{1}{\tau} \ln\left(\frac{T}{\mathcal{N}(t)}\right)\,. \tag{14}$$

If one then solves

$$\dot{\mathcal{N}} = -a\mathcal{N} = \mathcal{I}^{hom} + \mathrm{trace}(\mu \times \mathcal{I}^{\mu}) \tag{15}$$

for μ, with a given by (14), the value of the norm $\mathcal{N}(t + \tau)$ should be T, independent of its initial value. Therefore, (15) serves as a guide to formulate a practical method to choose free parameters in the equations with which the numerical solution behaves well with respect to the satisfaction of the constraints. Naturally, if one deals, as it is often the case, with more than one free parameter, (15) must be augmented with other conditions to yield a unique solution. This extra freedom is actually very useful in preventing large time-variations in the parameters that are sometimes needed in order to keep the constraints under control. These large variations do not represent a fundamental problem but a practical one, due to the small time stepping that they require in order to keep errors due to time integration reasonably small. One way to prevent this is by using this extra freedom to pick up the point in parameter space that not only gives the desired constraint growth, but also minimizes the change of parameters between two consecutive timesteps.

Rather than including the full details on the particular way we have implemented the method, we describe here a simple example to illustrate its application. Assume, for instance, that within a particular formulation only two free functions, $\{\kappa, \omega\}$, are employed. Equation (15) formally evaluates to

[1] There is a slight abuse of notation here, in the sense that a does not denote an index, as before. Similarly, the subscript in n_a indicates that the quantity is related to a through (14).

$$\dot{\mathcal{N}} = -a\mathcal{N} = \mathcal{I}^{hom} + \kappa\mathcal{I}^{\kappa} + \omega\mathcal{I}^{\omega} \ . \tag{16}$$

Now, we exploit the freedom in the free functions to adjust the rate of change of the energy $\mathcal{N}$ if the values of $\{\mathcal{I}^{hom}, \mathcal{I}^{\kappa}, \mathcal{I}^{\omega}\}$ are known. In practice, these are easily obtained during evolution. Once these are known, (16), coupled to the requirement that $\{\kappa, \omega\}$ vary as little as possible from one evaluation to another, results in a straightforward strategy to evaluate preferred values of the free parameters. This is done at a single resolution "test" run and, through interpolation in time, continuum, a priori defined parameters which keep the constraints under control for the given problem are obtained. Depending on the formulation of the equations, the free parameters might have to satisfy some conditions in order for symmetric hyperbolicity to hold, which can restrict the range of values these parameters can take. Nevertheless, even within a restricted window, the technique allows one to adopt the most convenient values these parameters should have for the problem at hand.

3 Applications

We now present applications of the techniques previously discussed. The goal is to illustrate how well-resolved simulations can indeed serve as a powerful tool to understand particular problems. To this end we have chosen a problem found in higher dimensional general relativity. A second application is that of the simulation of single black hole space-times, where the issue of the a priori lack of a preferred formulation is illustrated.

3.1 Bubble Space-Times

As a first application we concentrate on the study of *bubble space-times* and elucidate the dynamical behavior of configurations with both positive and negative masses and their possible connection to naked singularities. Bubble space-times have been studied extensively within five-dimensional Kaluza-Klein theory. These are five-dimensional space-times in which the circumference of the "extra" dimensions shrinks to zero on some compact surface referred to as the "bubble". These bubbles were initially studied by their relevance in the quantum instability of flat space-time [25], as bubbles can be obtained via semi-classical tunneling from it. They were later extended to include data corresponding to negative energy configurations (at a moment of time symmetry) [26, 27]. As mentioned, among the reasons for considering negative energy solutions is that naked singularities are associated with them. Therefore, these solutions are attractive tests of the cosmic censorship conjecture. Additionally, bubble space-times can also be obtained by double-Wick rotation of black strings, whose stability properties (or lack thereof) have been the subject of intense scrutiny in recent years. These features make bubble space-times both interesting and relevant for gravity beyond

four-dimensions, and thus attention has been devoted to fully understand their behavior. As we will see, even when the "analytical" study of the problem is greatly simplified by symmetry assumptions, many lingering questions remain and numerical simulations provided a viable way to shed light into them. Furthermore, these simulations were also key to 'digging out' a few unexpected features of the solution.

In order to obtain a complete description of the dynamical behavior of these space-times, a numerical code, implementing Einstein equations in 5D settings, and capable of handling the possibly strong curvature associated need be constructed. Fortunately, the assumption of a $\mathbf{SO(3)} \times \mathbf{U(1)}$ symmetry simplifies the treatment of the problem, which can be reduced to a $1+1$ manifold. This, in turn, renders the problem quite tractable by the currently available computational resources, though as we will see, considerable care must be placed at both analytical and numerical levels for an accurate treatment of the problem.

Initial Data. We consider a generalization of the time symmetric family of initial data presented in [27]. We start with a space-time endowed with the metric

$$ds^2 = -dt^2 + U(r)dz^2 + \frac{dr^2}{U(r)} + r^2 d\Omega^2 \ , \tag{17}$$

where $d\Omega^2 = d\vartheta^2 + \sin^2\vartheta d\varphi^2$ is the standard metric on the unit two-sphere S^2 and $U(r)$ is a smooth function that has a regular zero at some $r = r_+ > 0$, is everywhere positive for $r > r_+$ and converges to one as $r \to \infty$. The coordinate z parameterizes the extra dimension S^1 which has the period $P = 4\pi/U'(r_+)$. The resulting space-time $\{t, z, r \geq r_+, \vartheta, \varphi\}$ constitutes a regular manifold with the topology $R \times R^2 \times S^2$. The bubble is located where the circumference of the extra dimension shrinks to zero, that is, at $r = r_+$.

Additionally, we consider the presence of an electromagnetic field of the form

$$\frac{1}{2} F_{\mu\nu}\, dx^\mu \wedge dx^\nu = d\gamma(r) \wedge dz \ , \tag{18}$$

where $\gamma(r)$ is a smooth function of r that converges to zero as $r \to \infty$. The symmetries of the problem would also allow for a non-trivial electric component of the field. However, it is not difficult to show that Maxwell's equations imply that such a field necessarily diverges at the location of the bubble. For this reason, in the following, we only consider the case of vanishing electric field.

In this article, we consider initial data with

$$\gamma(r) = k(r_+^{-n} - r^{-n}) \ , \tag{19}$$

where k is an arbitrary constant and n an integer greater than one. This field generalizes the ansatz considered in [27], where only the case $n = 2$ was discussed, and allows for different interesting initial configurations. In

the time-symmetric case, initial data satisfying the Hamiltonian constraint obeys

$$U(r) = 1 - \frac{m}{r} + \frac{b}{r^2} - \frac{\tilde{k}^2}{r^{2n}} \,, \tag{20}$$

with $\tilde{k} \equiv kn/\sqrt{(n-1)(2n-1)}$ and free integration constants m and b. Here, the parameter m is related to the ADM mass via $M_{ADM} = m/4$. The fact that the bubble be located at $r = r_+$ requires that $0 = U(r_+) = 1 - \bar{m} + \bar{b} - \bar{k}^2$, where $\bar{m} \equiv m/r_+$, $\bar{b} \equiv b/r_+^2$, $\bar{k} \equiv \tilde{k}/r_+^n$. We also require

$$0 < r_+ U'(r_+) = 2 - \bar{m} + 2(n-1)\bar{k}^2 \tag{21}$$

and avoid the conical singularity at $r = r_+$ by fixing the period of z to $P = 4\pi/U'(r_+)$. It can be shown that the initial acceleration of the bubble area A with respect to proper time is given by

$$\ddot{A} = 8\pi \left[1 - \bar{m} - \frac{4\bar{k}^2}{3}(n-1)(n-2) \right] \,. \tag{22}$$

For $n = 2$, as discussed in [28], this implies that negative mass bubbles start out expanding (the initial velocity of the area is zero since we only consider time-symmetric initial data), while for large enough positive mass the bubble starts out collapsing. In the vacuum case, our numerical simulations suggest that initially collapsing bubbles undergo complete collapse and form a black string. In the non-vacuum case however, the strength of the electromagnetic field can modify this behavior completely. We will see that for small enough k the bubble continues to collapse whereas when k is large the bubble area bounces back and expands. Interesting behavior is obtained at the critical value for k which divides the phase space between collapsing and expanding solutions.

For $n > 2$ it is possible to obtain initial configurations with negative mass and negative initial acceleration [29]. This can potentially give rise to a collapsing bubble of negative energy, and thus to a naked singularity. However, our numerical results [29] suggest that cosmic censorship is valid: The bubble bounces back and starts out expanding.

Equations. In order to study the time evolution of the initial data sets given on a $t = \text{const.}$ slice of the metric (17) and the electromagnetic field (18), it is convenient to introduce a new radial coordinate $R = R(r)$ which facilitates the specification of regularity conditions at the bubble location. This new coordinate is defined by

$$R(r) = \sqrt{r^2 - r_+^2} \,, \qquad r > r_+ \,. \tag{23}$$

The metric (17) now reads

$$ds^2 = -\alpha^2 dt^2 + e^{2a} dR^2 + \frac{R^2}{r_+^2 + R^2} e^{2b} dz^2 + (r_+^2 + R^2) e^{2c} d\Omega^2 \,, \tag{24}$$

with $\alpha = 1$, $e^{-2a} = e^{2b} = (r_+^2 + R^2)U(R)/R^2$, $c = 0$. Since $U(R) = \text{const.} \cdot (R/r_+)^2 + O(R^4)$ near $R = 0$, and $U(R)$ converges to one in the asymptotic region, a and b are regular functions. An explicit example is the initial data corresponding to the zero mass Witten bubble [25] where $U = 1 - (r_+/r)^2$ and thus $a = b = 0$. When studying the time evolution of the initial data sets discussed above, we consider the metric (24) where α, a, b and c are functions of t and R. As we will see, the coordinate R is well suited for imposing regularity conditions at the bubble location since (R, z) represent polar coordinates near the bubble, $R = 0$ being the center, and z assuming the role of the angular coordinate. In order to avoid a conical singularity, z must have the period $2\pi r_+ e^{a-b}$. For this to be constant we need to impose the boundary condition $a(t, 0) - b(t, 0) = \text{const.}$ at $R = 0$.

Similarly, the electromagnetic field (18) is written in the form

$$\frac{1}{2} F_{\mu\nu}\, dx^\mu \wedge dx^\nu = \frac{R}{\sqrt{r_+^2 + R^2}} e^b \left(\pi_\gamma dt + d_\gamma dR\right) \wedge dz \,, \tag{25}$$

where the functions π_γ and d_γ depend on t and R and satisfy $\pi_\gamma = 0$ and $d_\gamma = e^{-b}\partial_r \gamma$ at the initial time.

We choose the following gauge condition for the lapse

$$\log(\alpha) = a + \lambda(b + 2c) \,, \tag{26}$$

with a parameter λ which, in our simulations, is either zero or one. For $\lambda = 1$ the resulting gauge condition is strongly related to the densitized lapse condition often encountered in hyperbolic formulations of Einstein's equations: Indeed, the square root of the determinant of the four metric belonging to (24) is given by $\sqrt{g^{(4)}} = e^{a+b+2c} R \sqrt{r_+^2 + R^2} \sin\vartheta$, so (26) sets α equal to the square root of the determinant of the four metric but divides the result by the factor $R\sqrt{r_+^2 + R^2} \sin\vartheta$ which is singular at the bubble, at the poles $\vartheta = 0, \pi/2$ and in the asymptotic region. For $\lambda = 0$, the condition (26) implies that the two-metric $-\alpha^2 dt^2 + e^{2a} dR^2$ is in the conformal flat gauge. As we will see, the principal part of the evolution equations is governed by the d'Alembertian with respect to this metric. Since the two-dimensional d'Alembertian operator is conformally covariant, the resulting equations are semi-linear in that case. In particular, this implies that the characteristic speeds do not depend on the solution that is being evolved.

The field equations resulting from the five-dimensional Einstein-Maxwell equations split into a set of evolution equations and a set of constraints. The evolution equations can be written as

$$\ddot{A} = e^{-\lambda F}\left[(A'+2G')e^{\lambda(2B+F)}\right]' - 3(\lambda-1)(C'+G')^2 e^{2\lambda B} - (\lambda+1)V$$
$$+2\lambda\dot{A}\dot{B} - \lambda(\lambda+1)\dot{B}^2 - 3(\lambda+1)\dot{C}^2 + G\left[(1-\lambda)\pi_\gamma^2 - (1+\lambda)e^{2\lambda B}d_\gamma^2\right] , \quad (27)$$

$$\ddot{B} = e^{(\lambda-1)B-(\lambda+2)F}\left[B' e^{(\lambda+1)B+(\lambda+2)F}\right]' + \frac{3r_+^2+2R^2}{(r_+^2+R^2)^2}\, e^{2\lambda B}$$
$$+ (\lambda-1)\dot{B}^2 - 2V , \quad (28)$$

$$\ddot{C} = e^{(\lambda-1)B-F}\left[(C'+G')e^{(\lambda+1)B+F}\right]' - V + (\lambda-1)\dot{B}\dot{C},$$
$$+ \frac{2G}{3}\left[\pi_\gamma^2 - e^{2\lambda B}d_\gamma^2\right] , \quad (29)$$

$$\dot{\pi}_\gamma = e^{\lambda B-2(C+G)}\left[d_\gamma e^{\lambda B+2(C+G)}\right]' + (\lambda\dot{B}-2\dot{C})\pi_\gamma , \quad (30)$$

$$\dot{d}_\gamma = \frac{\sqrt{r_+^2+R^2}}{R} e^{-(B-2C)}\left[\pi_\gamma \frac{R}{\sqrt{r_+^2+R^2}} e^{B-2C}\right]' - (\dot{B}-2\dot{C})d_\gamma , \quad (31)$$

where we have set $A = a + \lambda b + 2(\lambda+1)c$, $B = b + 2c$, $C = c$ and $G = \log(r_+^2+R^2)/2$, $F = \log(R) + G$ and $V = e^{2(A-3C)}/(r_+^2+R^2)$. Here, and in the following, a dot and a prime denote differentiation with respect to t and R, respectively. The evolution equations constitute a hyperbolic system on the domain $R > 0$.

The constraints are the Hamiltonian and the R component of the momentum constraint, given by $\mathcal{C} = 0$, $\mathcal{C}_R = 0$, where

$$\mathcal{C} = e^{(\lambda-1)B-(\lambda+2)F}\left[e^{(\lambda+1)B+(\lambda+2)F}B'\right]'$$
$$+ \left[\frac{3r_+^2+2R^2}{(r_+^2+R^2)^2} - (B'+F')(A'+2G') + 3(C'+G')^2\right]e^{2\lambda B}$$
$$- V - (\dot{A}-\lambda\dot{B})\dot{B} + 3\dot{C}^2 + G\left[\pi_\gamma^2 + e^{2\lambda B}d_\gamma^2\right] , \quad (32)$$

$$\mathcal{C}_R = e^{A-2C}\left[e^{-(A-2C)}\dot{B}\right]' - (B'+F')\left[\dot{A} - (\lambda+1)\dot{B}\right]$$
$$+ 2(C'+G')(3\dot{C}-\dot{B}) + 2G\pi_\gamma d_\gamma . \quad (33)$$

Regularity Conditions. The evolution equations contain terms proportional to e^{-F} which diverge like $1/R$ near $R = 0$, and therefore, regularity conditions have to be imposed at $R = 0$. This is achieved by demanding the boundary conditions

$$A' = B' = C' = \pi_\gamma = 0 \quad \text{at } R = 0 . \quad (34)$$

Assuming that the fields are smooth enough near $R = 0$, it then follows that the right-hand side of the evolution equations is bounded for $R \to 0$. Next, as discussed above, the avoidance of a conical singularity at $R = 0$ requires that $A - (\lambda+1)B = a - b$ must be constant at $R = 0$. We show that this condition is a consequence of the evolution and constraint equations, and

of the regularity conditions (34). Using the evolution equations in the limit $R \to 0$ and taking into account the conditions (34), we find

$$\partial_t \left\{ e^{(1-\lambda)B} \left[\dot{A} - (\lambda+1)\dot{B} \right] \right\} \Big|_{R=0} = -(\lambda+1) e^{(1-\lambda)B} \mathcal{C} \Big|_{R=0} . \tag{35}$$

This means that if the Hamiltonian constraint is satisfied at $R = 0$ (or in the case that $\lambda = -1$ even if the constraints are violated), the condition $A - (\lambda+1)B|_{R=0} = \text{const.}$ will hold provided that the initial data satisfies $\dot{A} - (\lambda+1)\dot{B}|_{R=0} = 0$. Next, we analyze the propagation of the constraint variables $\mathcal{C}$ and $\mathcal{C}_R$ and show that the regularity conditions (34) and the evolution equations imply that the constraints are satisfied at each time provided they are satisfied initially.

Propagation of the Constraints. First, we notice that the vanishing of the momentum constraint requires that $\dot{A} - (\lambda+1)\dot{B}|_{R=0} = 0$ because of the factor F' which diverges like $1/R$ near $R = 0$ in the definition of $\mathcal{C}_R$. This is precisely the condition $a(t,0) - b(t,0) = \text{const.}$ discussed above. However, for this condition to hold, we first have to show that the momentum constraint actually vanishes. In order to see this, we regularize the constraint variables and define $\tilde{\mathcal{C}} = e^F \mathcal{C}$, $\tilde{\mathcal{C}}_R = e^F \mathcal{C}_R$. Now the regularity conditions (34) imply that $\tilde{\mathcal{C}}_R$ is regular and that $\tilde{\mathcal{C}}$ vanishes at $R = 0$. As a consequence of the evolution equations and Bianchi's identities, the constraint variables obey the following evolution system

$$\partial_t \tilde{\mathcal{C}} = e^{(\lambda-1)B} \partial_R \left[e^{(\lambda+1)B} \tilde{\mathcal{C}}_R \right] + (3\lambda - 1)\dot{B}\tilde{\mathcal{C}} , \tag{36}$$

$$\partial_t \tilde{\mathcal{C}}_R = e^{-(\lambda+1)B - \lambda F} \partial_R \left[e^{(\lambda+1)B + \lambda F} \tilde{\mathcal{C}} \right] + (\lambda - 1)\dot{B}\tilde{\mathcal{C}}_R \tag{37}$$

which is regular at $R = 0$. Defining the energy norm

$$\mathcal{E}(t) = \frac{1}{2} \int_0^\infty \left(e^{2B + \lambda F} \tilde{\mathcal{C}}^2 + e^{2(\lambda+1)B + \lambda F} \tilde{\mathcal{C}}_R^2 \right) dR , \tag{38}$$

taking a time derivative and using the equations (36), (37) we obtain

$$\frac{d}{dt} \mathcal{E} = e^{2(\lambda+1)B + \lambda F} \tilde{\mathcal{C}} \tilde{\mathcal{C}}_R \Big|_0^\infty + \lambda \int_0^\infty \dot{B} \left(3 e^{2B + \lambda F} \tilde{\mathcal{C}}^2 + 2 e^{2(\lambda+1)B + \lambda F} \tilde{\mathcal{C}}_R^2 \right) dR . \tag{39}$$

The boundary term vanishes because of the regularity conditions at $R = 0$ and under the assumption that all fields fall off sufficiently fast as $R \to \infty$. If $\dot{B}$ is smooth and bounded, we can estimate the integral on the right-hand side by a constant C times $\mathcal{E}$, and it follows that $\mathcal{E}(t) \leq e^{Ct} \mathcal{E}(0)$. This shows that if the constraints are satisfied initially, they are also satisfied for all $t > 0$ for which a smooth solution exists. In the gauge where $\lambda = 0$ we even obtain the result that the norm of the constraints cannot grow in time.

To summarize, the boundary conditions (34) imply that the constraints $\mathcal{C} = 0$, $\mathcal{C}_R = 0$ and $\dot{A} - (\lambda+1)\dot{B} \Big|_{R=0} = 0$ are preserved throughout evolution.

Outer Boundary Conditions. For numerical computations, our domain extends from $R = 0$ to $R = R_{max}$ for some $R_{max} > 0$. Now we have to replace the estimate (39) by the estimate

$$\frac{d}{dt}\,\mathcal{E} = e^{2(\lambda+1)B+\lambda F+A_0}\tilde{\mathcal{C}}\tilde{\mathcal{C}}_R\Big|_0^{R_{max}} + C\mathcal{E}\ , \tag{40}$$

and it only follows that the constraints are zero if we control the boundary term at $R = R_{max}$. For this reason, we impose the momentum constraint, $\mathcal{C}_R = 0$, at $R = R_{max}$. This condition results in an evolution equation for B' at the outer boundary. We combine this condition with the Sommerfeld-like conditions at $R = R_{max}$,

$$\dot{A} + A' = 0, \qquad \dot{C} + C' = 0, \qquad \pi_\gamma + d_\gamma = 0\ . \tag{41}$$

Numerical Implementation. Next, we discuss the numerical implementation of the above constrained evolution system. In order to apply the discretization techniques discussed in Sect. 2 we first recast the evolution equations into first order symmetric hyperbolic form by introducing the new variables $\pi_A = \dot{A}$, $\pi_B = \dot{B}$, $\pi_C = \dot{C}$ and $d_A = A' + 2G'$, $d_B = B'$, $d_C = C' + G'$. The resulting first order system is then discretized by the method of lines. Let us first discuss the spatial discretization which requires special care at $R = 0$ because of the coefficients proportional to $1/R$ that appear in the evolution equations. To this end, consider the following family of toy models

$$\dot{\pi} = R^{1-n}\partial_R(R^{n-1}d), \tag{42}$$

$$\dot{d} = \partial_R\pi\ , \tag{43}$$

where $R > 0$ is the radial coordinate, and $n = 1, 2, 3, ...$ We impose the regularity condition $d = 0$ at $R = 0$, which, for sufficiently smooth fields, implies that $\dot{\pi} = n\partial_R d$ at $R = 0$, and assume that the fields vanish for R sufficiently large. The toy model (42–43) corresponds to the n-dimensional wave equation for spherically symmetric solutions. The principal part of our evolution system has precisely this form near $R = 0$, where n is given by $\lambda+1$, $\lambda + 3$, 2, 1 for the evolution equations for π_A, π_B, π_C and π_γ, respectively. The system (42–43) admits the conserved energy

$$E = \frac{1}{2}\int_0^\infty R^{n-1}\left(\pi^2 + d^2\right) dR\ . \tag{44}$$

A second order accurate and stable numerical discretization of the system (42–43) can be obtained as follows: We assume a uniform grid $R_j = j\Delta R$, $j = 0, 1, 2...$, approximate the fields π and d by grid functions $\pi_j = \pi(R = R_j)$, $d_j = d(R = R_j)$, and consider the semi-discrete system

$$\dot{\pi}_j = R_j^{1-n}D_0(R^{n-1}d)_j \quad \text{for } j > 0 \text{ and } \quad \dot{\pi}_0 = \frac{n}{\Delta R}\,d_1\ , \tag{45}$$

$$\dot{d}_j = D_0\pi_j \quad \text{for } j > 0 \text{ and } \quad \dot{d}_0 = 0\ , \tag{46}$$

where for a grid function u_j, $(D_0 u)_j = (u_{j+1} - u_{j-1})/(2\Delta R)$ is the second order accurate centered differencing operator. It is not difficult to check that this scheme preserves the discrete energy

$$E_{\text{discrete}} = \frac{\Delta R}{2} \sum_{j=1}^{\infty} R_j^{n-1} \left(\pi_j^2 + d_j^2\right) + \frac{\Delta R}{4n} R_1^{n-1} \pi_0^2 \tag{47}$$

which proves the numerical stability of the semi-discrete system. Finally, we use a third order Runge–Kutta algorithm in order to perform the integration in time. By a theorem of Levermore [9], this guarantees the numerical stability of the fully discrete system for small enough Courant factor.

We apply these techniques for the discretization of our coupled system. The outer boundary conditions are implemented by a projection method. Of course, the resulting system is much more complicated than the simple toy model problem presented above, and we have no a priori proof of numerical stability. Nevertheless, we find the above analysis useful as a guide for constructing the discretization. Our resulting code is tested by running several convergence tests, and its accuracy is tested by monitoring the constraint variables $\mathcal{C}$ and $\mathcal{C}_R$ and the quantity $\dot{A} - (\lambda + 1)\dot{B}\Big|_{R=0} = 0$.

Results. Here we discuss the results for the numerical evolution of the initial data defined by (17–20). We start by reviewing the evolution of the initially expanding bubbles and the initially collapsing negative mass bubbles [29] and then focus on the initially collapsing positive mass bubbles.

Brill-Horowitz Initially Expanding Case. The Brill-Horowitz initial data ($n = 2$) in the case of vanishing electromagnetic field is evolved. The bubble area A as a function of the proper time τ at the bubble is shown in Fig. 1 for different values of the mass parameter m. As expected, the lower the mass of the initial configuration, the faster the expansion. Empirically, and for the parameter ranges used in our runs, we found that at late times the expansion rate obeys

$$\frac{\dot{A}}{A} \approx \frac{2 - \bar{m}}{r_+(\tau = 0)} \,, \tag{48}$$

where a dot denotes the derivative with respect to proper time τ. In particular this approximation is valid for the bubble solution exhibited by Witten [25] which describes the time evolution in the case $\bar{m} = 0$.

Collapsing Negative Mass Case. We here restrict to cases with negative masses that start out collapsing. Interestingly enough we find that even when starting with large initial negative accelerations, which in turn make the bubble shrink in size to very tiny values, it bounces back without ever collapsing into a naked singularity. As an example, Fig. 2 shows the bubble's area versus time for different values of n and k. The initially collapsing bubbles decrease

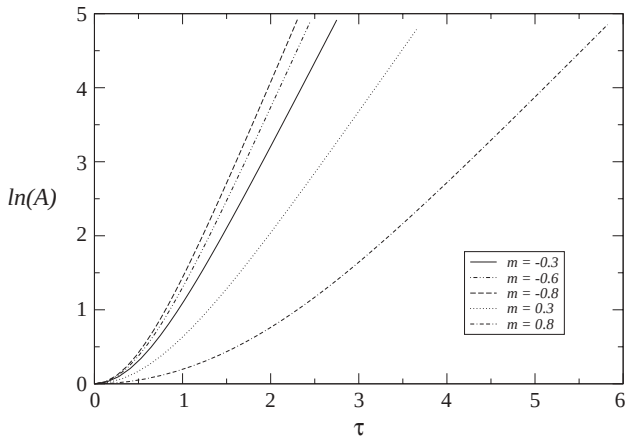

Fig. 1. Bubble area vs. proper time at the bubble. In this and the following plots, we set $r_+ = 1$. The figure shows five illustrative examples of bubbles whose initial acceleration is positive. As it is evident, the expansion of the bubble continues and the difference is the rate of the exponential expansion. The relative error in these curves, estimated from the appropriate Richardson extrapolated solution in the limit $\Delta \to 0$, is well below 0.001%

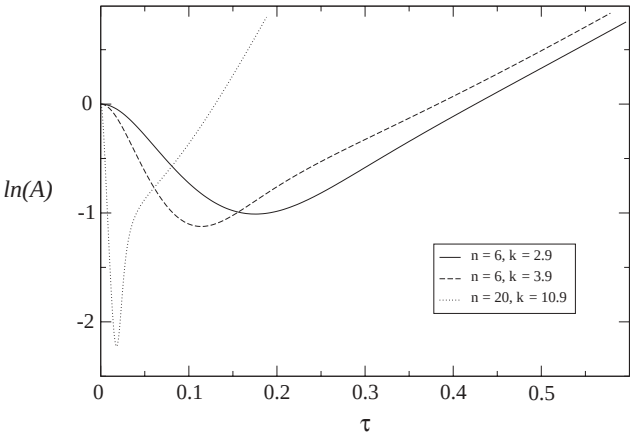

Fig. 2. Bubble area vs. proper time at the bubble. The figure shows three illustrative examples of bubbles with negative mass ($m = -0.1$ each) whose initial acceleration is negative. As it is evident, the collapse of the bubble is halted and the trend is completely reversed. The error in these curves is estimated to be well below 0.001%

in size in a noticeable way but this trend is halted and the bubbles bounce back and expand. Although we have not found a simple law as that in (48), clearly the bubbles expand exponentially fast. Therefore, it seems not to be possible to "destroy" the bubble and create a naked singularity. This situation

is somewhat similar to the scenarios where one tries to "destroy" an extremal Reissner-Nordström black hole by attempting to drop into it a test particle with high charge to mass ratio. There, the electrostatic repulsion prevents the particle from entering the hole [30].

Brill-Horowitz Initially Collapsing Case. Next, we analyze the Brill-Horowitz initial data for the case in which the bubble is initially collapsing (notice that for $n = 2$ this implies that the ADM mass is positive). While our numerical simulations reveal that in the absence of the gauge field such a bubble continues to collapse, we also show that when the gauge field is strong enough, the bubble shrinks at a rate which decreases with time and then bounces back.

Obviously, if the collapse trend were not halted, a singularity should form at the origin. Since the ADM mass is positive, one expects this singularity to be hidden behind an event horizon, and one should obtain a black string. In fact, for the solutions which are initially collapsing and which have vanishing gauge field, we observe the formation of an apparent horizon. Furthermore, we compute the curvature invariant quantity Ir_{AH}^4 at the apparent horizon (as discussed in [31]), where $I = R_{abcd}R^{abcd}$ is the Kretschmann invariant and r_{AH} the areal radius of the horizon. For a neutral black string, this invariant is 12. Figure 3 shows how this value is attained after the apparent horizon forms for representative vacuum cases (with $m = 1.1$ and $m = 1.99$) this, together with the formation of apparent horizons, provides strong evidence for the formation of a black string.

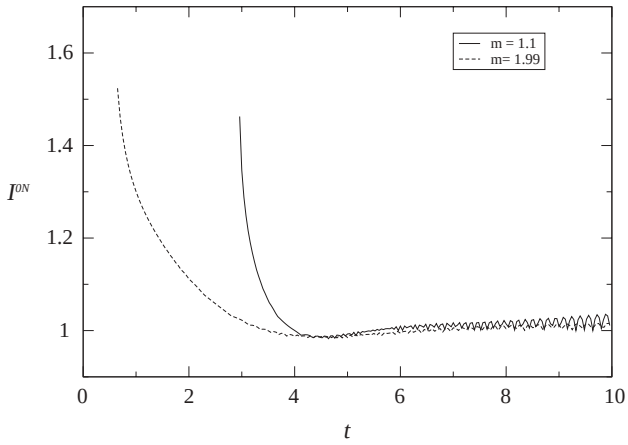

Fig. 3. Rescaled Kretschmann invariant $I^{0N} \equiv Ir_{AH}^4/12$ vs. asymptotic time for $m = 1.1$ (*solid line*) and 1.99 (*dashed line*). The first non-zero values of the lines mark the formation of the apparent horizon. After some transient period, both lines approach the value of 1 suggesting a black string has formed

As mentioned, for strong enough gauge fields, the previously described dynamics is severely affected. Fig. 4 (left panel) shows the bubble area vs. proper time for different values of k. For large values the bubble "bounces" back and expands while for small ones the bubble collapses. There is a natural transition region separating these two possibilities. Tuning the value of k one can reveal an associated critical phenomenon, the 'critical solution' being a member of the family of static solutions given by

$$ds^2 = -V(r)dt^2 + \frac{V(r)}{U(r)}dr^2 + \frac{U(r)}{V(r)^2}dz^2 + r^2V(r)d\Omega^2, \tag{49}$$

$$\frac{1}{2}F_{\mu\nu}\,dx^\mu \wedge dx^\nu = \pm\frac{1}{2}\sqrt{3r_-(r_+ - r_-)}\,\frac{dr \wedge dz}{r^2V(r)^2}\ , \tag{50}$$

where $V(r) = 1 - r_-/r$ and $U(r) = 1 - r_+/r$. The parameters r_- and r_+ ($> r_-$) are related to the period of the z coordinate and to the ADM mass via $P = 4\pi r_+(1 - r_-/r_+)^{3/2}$ and $M_{ADM} = r_+/4$. Since the quantities P and M_{ADM} are conserved, the member of the family of static solutions the dynamical solution approaches can be determined a priori from the initial data.

Figure 4 (right panel) displays the time T defined as the length of asymptotic time during which the bubble's area stays within 1% of the minimum value attained when the bubble bounces back. This is a measure of how long the solution stays close to the static solution as a function of the parameter k. Empirically, we find the law

$$T = -r_+\Gamma \log|k - k_c| + T_1\ , \tag{51}$$

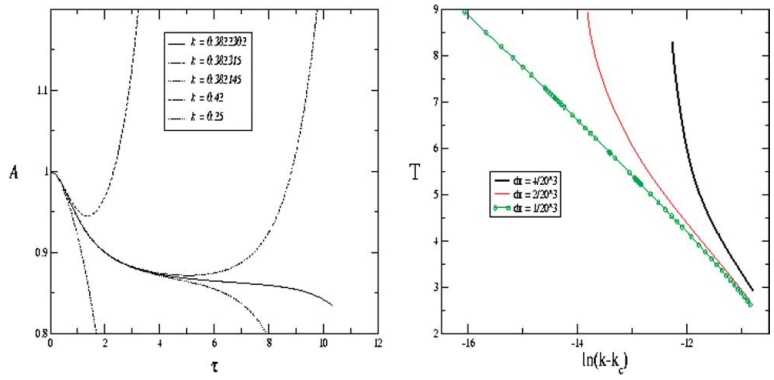

Fig. 4. *Left Panel.* Area values vs. proper time at the bubble for different values of k and $m = 1.1$. By tuning the value of k appropriately, the amount of time that the area remains fairly constant can be extended for as long as desired. *Right Panel.* The time T which is a measure of how long the solution stays close to the static solution vs. the logarithm of the difference between the parameter k and its critical value. A linear interpolation gives the value $\gamma \approx 1.2$

with a parameter $\Gamma \approx 1.2$ that does not seem to depend on the family of initial data chosen. This universality property is reinforced by the linear stability analysis of the critical solutions (49,50) performed in [32] where we prove that each solution has precisely one unstable linear mode growing like $\exp(\Omega t/r_+)$ with a universal Lyapunov exponent of $\Omega \approx 0.876$. This explains the law (51) with $\Gamma = 1/\Omega \approx 1.142$.

3.2 Black Holes

As one of the applications that we have chosen to illustrate the use of the techniques previously discussed we consider here the evolution of single non-spinning black holes. Even when the data provided correspond to spherically symmetric and vacuum scenarios, as we will see, obtaining a long term stable implementation is not a trivial task. For additional information, and a more general treatment, we refer the reader to [33].

Formulation. We adopt the symmetric hyperbolic family of formulations introduced in [34]. This is a first order formulation whose evolved variables are given by $\{g_{ij}, K_{ij}, d_{kij}, \alpha, A_i\}$ with g_{ij} the induced metric on surfaces at $t = \text{const.}$, K_{ij} the extrinsic curvature, d_{kij} are first derivatives of the metric, $d_{kij} = \partial_k g_{ij}$, α is the lapse, and A_i are normalized first derivatives of the lapse, $A_i = \alpha^{-1}\partial_i \alpha$.

The Einstein equations written in this formulation are subject to the physical constraints, the Hamiltonian and momentum constraints, as well as non-physical constraints, which arise from the variable definitions. The non-physical constraints are

$$C_{A_i} = A_i - N^{-1}\partial_i N = 0 \ , \ \ C_{kij} = d_{kij} - \partial_k g_{ij} = 0 \ , \ \ C_{lkij} = \partial_{[l} d_{k]ij} = 0 \ . \tag{52}$$

The constraints are added to the field equations and the space-time *constraint-functions* $\{\gamma(t), \zeta(t), \eta(t), \chi(t), \xi(t)\}$ are introduced as multiplicative factors to the constraints. While these quantities are sometimes introduced as parameters, we extend them to time-dependent functions. For simplicity in this work, we set $\zeta = -1$. Requiring that the evolution system is symmetric hyperbolic imposes algebraic conditions on these factors, and they are not all independent. If we require that all the characteristic speeds are "physical" (i.e. either normal to the spatial hypersurfaces or along the light cone), then we obtain two symmetric hyperbolic families. One family has a single free parameter, $\chi(t)$,

$$\text{Single constraint-function system} \begin{cases} \gamma = -\frac{1}{2} \\ \zeta = -1 \\ \eta = 2 \\ \xi = -\frac{\chi}{2} \\ \chi \neq 0 \end{cases} \tag{53}$$

and another symmetric system with two varying constraint-functions $\{\eta(t), \gamma(t) \neq -1/2\}$:

$$\text{Two constraint-function system} \begin{cases} \zeta = & -1 \\ \chi = & -\frac{\gamma(2-\eta)}{1+2\gamma} \\ \xi = & -\frac{\chi}{2} + \eta - 2 \\ \gamma \neq & -\frac{1}{2} \\ \eta & \end{cases} \tag{54}$$

Initial Data and Boundaries. Initial data for a Schwarzschild black hole are given in ingoing Eddington-Finkelstein coordinates. The shift β^i will be considered an a priori given vector field while the lapse is evolved to correspond to the time harmonic gauge with a given source function. This gauge source function is taken from the exact solution, such that in the high-resolution limit $\alpha = (1 + 2M/r)^{-1/2}$.

Black hole excision is usually based on the assumption that an inner boundary (IB) can be placed on the domain such that information from this boundary does not enter the computational domain. This requirement places strenuous demands on cubical excision for a Schwarzschild black hole in Kerr-Schild, Painlevé–Gullstrand or the Martel–Poisson [35] coordinates: the cube must be inside $0.37\,M$ in each direction. This forces one to excise very close to the singularity, where gradients in the solution can become very large, requiring very high resolution near the excision boundary to adequately resolve the solution. This requirement follows directly from the physical properties of the Schwarzschild solution in these coordinates, and is independent of the particular formulation of the Einstein equations [6].

With our current uniform Cartesian code, however, we do not have enough resolution to adequately resolve the Schwarzschild solution near the singularity. Thus, we place the inner boundary inside the event horizon, but outside the region where all characteristics are outgoing. The difference stencils are one-sided at the inner boundary, and no boundary conditions are explicitly applied. Testing various locations we find that placing the inner boundary at $1.1\,M$ gives reasonable results for the resolutions we are able to use, $\triangle x = \triangle y = \triangle z = M/5$, $M/10$, $M/20$. We are working to resolve this inconsistency in our code by using coordinate systems that conform to the horizon's geometry.

We performed numerical experiments with the outer boundary at three different locations, $5\,M$, $10\,M$ and $15\,M$. Boundary conditions for the outer boundary are applied using the orthogonal projection technique referenced above, by "freezing" the incoming characteristic modes. That is, their time derivative is set to zero through an orthogonal projection. This makes use of the fact that one knows that the continuum, exact solution is actually stationary. While this would not be useful in the general case, as we shall see, even in such a simplified case the constraint manifold seems to be unstable. We are currently working on extending the boundary treatment to allow for

constraint-preserving boundary conditions and studying the well posedness of the associated initial-boundary value problem.

Having set up consistent initial and boundary data, in a second order accurate implementation using the techniques mentioned in Sect. 2, we now concentrate on simulating a stationary black hole space-time. As we will see below, even in this simple system, one encounters difficulties to evolve the system for long times. In particular, as has been illustrated in several occasions, the length of time during which a reliable numerical solution is obtained varies considerably depending on the values of the free parameters in the formulation. These parameters play no role at the constraint surface; however, off this constraint surface, these parameters have a sensible impact. Hence, at the numerical level – where generic data is only approximately at this surface -, it is necessary to adopt preferred values of these parameters. These, in turn, will depend not only on the physical situation under study but also on the details of the particular implementation (order of convergence, etc). As we argued in Sect. 2, the constraint minimization method provides a practical way to adopt these parameters. We next illustrate this in numerical simulations of Schwarzschild space-time.

Testing Constraint Minimization. We concentrate here on black hole simulations performed using the symmetric hyperbolic formulation with two constraint functions. The single function family and its disadvantages for constraint minimization are discussed in [33].

Black Hole Numerical Results. As a first attempt to numerically integrate the Einstein equations, one could simply fix the parameters η and γ to constant values. Lacking knowledge of preferred values for these parameters we might simply set $\eta = 0$ and $\gamma = 0$. Evolutions of the Schwarzschild space-time for these parameter choices, however, show that the solution is quickly corrupted, and the solution diverges. Figure 5 shows the error in the numerical solution with respect to the exact solution for three resolutions. While the code converges, the error at a single resolution grows without bound as a function of time.

We now apply the constraint minimization technique to evolutions of a Schwarzschild black hole. The constraint functions $\eta(t)$ and $\gamma(t)$ will now vary in time, and both will be used to control the constraint growth. With two functions we can attempt to minimize changes in the functions themselves. This is advantageous because smoothly varying functions seem to yield better numerical results. Thus, $\eta(t)$ and $\gamma(t)$ are chosen at time step $n+1$ to minimize the quantity

$$\triangle := [\eta(n+1) - \eta(n)]^2 + [\gamma(n+1) - \gamma(n)]^2 \ . \tag{55}$$

$\dot{\mathcal{N}}$ is nonlinear in γ but linear in η, allowing one to solve for η such that $\dot{\mathcal{N}} = -a\mathcal{N}$,

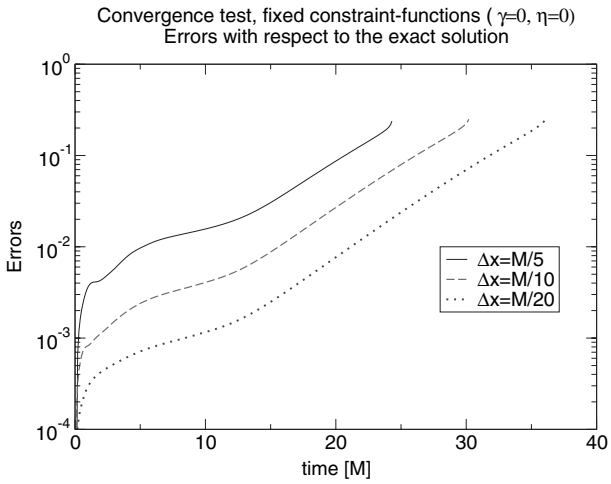

Fig. 5. Two-constraint-function family, with fixed values $\gamma = 0 = \eta$, inner and outer boundaries at $1.1\,M$ and $5\,M$, respectively

$$\eta = \frac{-(a\mathcal{N} + \mathcal{I}^{hom} + \mathcal{I}^{\gamma}\gamma)(1+2\gamma) + 2\gamma\mathcal{I}^{\chi}}{\mathcal{I}^{\eta}(1+2\gamma) + \gamma\mathcal{I}^{\chi}} \tag{56}$$

where, as in Sect. 3, a is given by (14). γ is chosen from some arbitrary, large interval. The corresponding η given by (56) is computed, and the pair (η, γ) that minimizes $\triangle$ defined in (55) is chosen. γ and η may be freely chosen, except that $\gamma \neq -1/2$, giving two "branches": γ always larger than $-1/2$, and γ always smaller than $-1/2$. We have only explored the $\gamma < -1/2$ branch using the seed values $\eta = 0$, $\gamma = -1$. In order to keep the variation of the parameters between two consecutive timesteps reasonably small, we have needed to set the tolerance value for the constraints energy roughly one order of magnitude larger than the initial discretization error, and n_a to either 10^2 or 10^3. This means that the constraints' energy, though in a longer timescale, will still grow.

The outer boundary is first placed at $5\,M$. Figure 6 shows the energy of the constraints and the error with respect to the exact solution. The corresponding constraint functions are shown in Fig. 7. The large variation in the functions near the end of the run appears to be a consequence of other growing errors. In Fig. 8 the minimization is stopped at $750\,M$, and the functions are fixed to $\eta = -1.88$, $\gamma = -1.00$ for the remainder of the run. The solution diverges at approximately the same time.

Another measure of the error in the solution is the mass of the apparent horizon, as shown in Fig. 9. After some time, the mass approximately settles down to a value that is around $1.009\,M$, which corresponds to an error of the order of one part in one thousand. For the higher resolution, the apparent

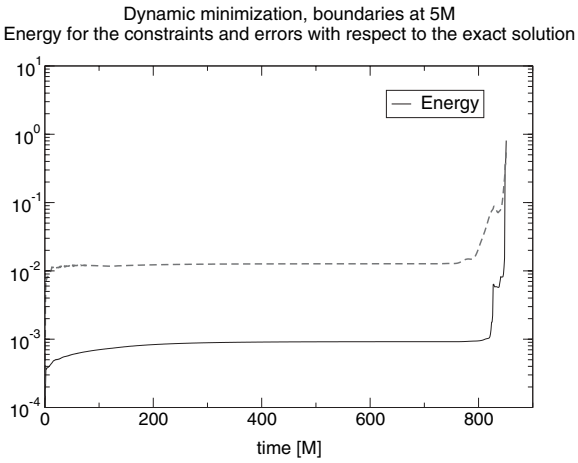

Fig. 6. This figure shows the constraint energy and the error with respect to the exact solution. Dynamic minimization is done with boundaries at $5\,M$, $\triangle x = M/5$, $T = 10^{-3}$, and $n_a = 10^3$

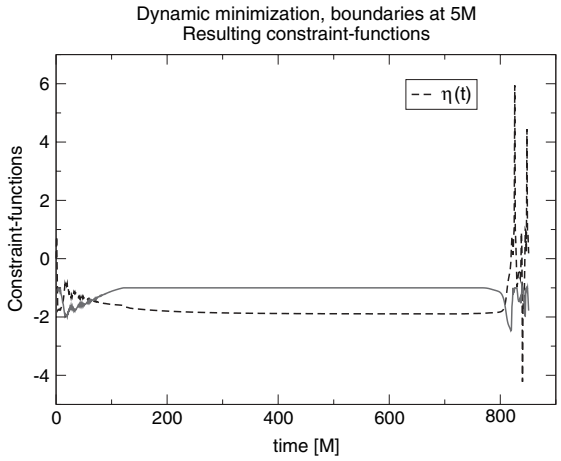

Fig. 7. This figure shows the constraint functions for the run described in Fig. 6

horizon mass at late times becomes indistinguishable from $1\,M$, given the expected level of discretization errors.

The outer boundary is now placed at $15\,M$. Figure 10 shows results for data equivalent to those discussed for Fig. 6. The initial discretization value for the energy is 7.6459×10^{-6}, and $T = 10^{-5}$, $n_a = 100$ was used. The minimization of the constraint-functions is stopped at $450\,M$, at which point the constraint-functions are approximately constant, and equal to

$$\eta = -1.35 \times 10^{-1} \quad , \quad \gamma = -3.39 \; . \tag{57}$$

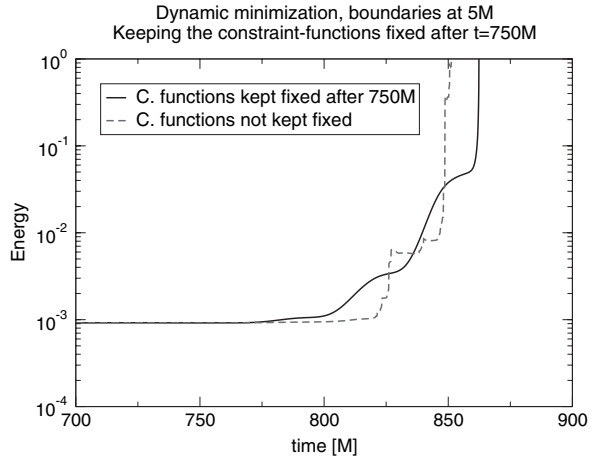

Fig. 8. Same as previous figure, but keeping the constraint-functions constant after $750\,M$. The figure compares the resulting energy for the constraints with that of the previous figure (shown at late times only, since because of the setup the runs are identical up to $t = 750\,M$)

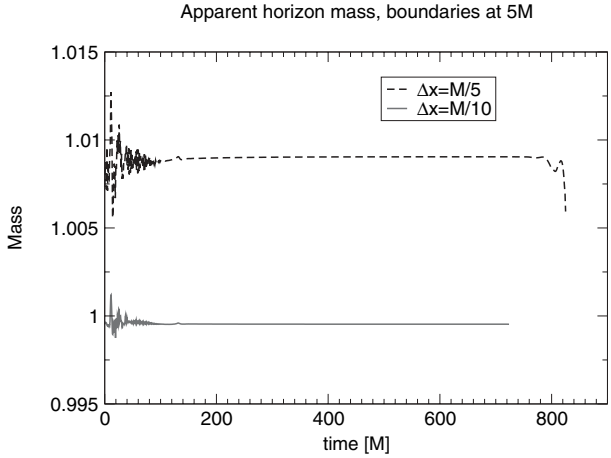

Fig. 9. This figure shows the black hole mass calculated from the apparent horizon with dynamic constraint-function values. The higher resolution simulation ran out of computing time. The apparent horizons were found using Thornburg's apparent horizon finder [36]

Figure 10 shows that the dependence of the lifetime on the location of the outer boundaries is not monotonic, as for this case the code runs for, roughly, $1000\,M$, while with boundaries at $10\,M$ and $5\,M$ it ran for around $700\,M$, and $800\,M$, respectively. A detailed analysis of such dependence would be computationally expensive and beyond the scope of this work, and may even

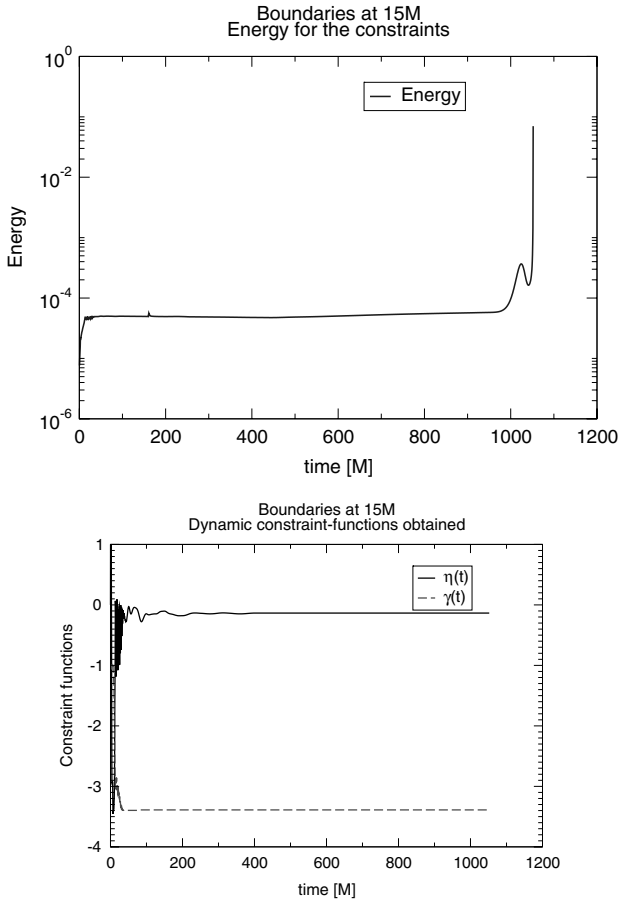

Fig. 10. Dynamic minimization done with boundaries at $15\,M$, $\triangle x = M/5$, $T = 10^{-5}$, and $n_a = 10^2$. The constraint-functions are constant for $t \geq 450\,M$, where they are $\eta = -0.135$, $\gamma = -3.389$. Thus, the constraint functions do not respond when the code is about to crash

depend on the details of the constraint minimization, such as the values for T and n_a. However, comparing Fig. 5 with Figs 6–9, we see that the constraint minimization considerably improves the lifetime of the simulation, as expected.

4 Final Words

We have chosen two problems to illustrate both the power of numerical simulations of Einstein's equations and some of the difficulties encountered in obtaining accurate numerical solutions. This is especially relevant for black

hole systems, where different poorly understood issues coupled to lack of sufficient computational power make it much more difficult to advance at a sustained pace towards the final goal of producing a reliable description of a binary black hole system. However, it is clear that that goal outweighs these difficulties. As the bubble problem illustrates, a robust implementation was not only key to responding to open questions but also proved to be the way to observing other phenomena not previously considered. Not only did it show that a priori possible way to violate cosmic censorship is invalid, but it also revealed the existence of critical phenomena, which, in turn, can be used to shed further light on the stability of black string systems [32].

Fortunately, a substantial body of work in recent years has begun to address a number of these questions. A better understanding of the initial boundary-value problem in general relativity, advances in the definition of initial data and gauge choices coupled to several modern numerical techniques are having a direct impact on current numerical efforts. It seems reasonable to speculate that if this trend continues, the ultimate goal will be within reach in a not-too-distant future.

Acknowledgements

This work was supported in part by the NSF under Grants No: PHY0244335, PHY0326311, INT0204937 to Louisiana State University, the Research Corporation, the Horace Hearne Jr. Institute for Theoretical Physics, NSF Grant No. PHY-0099568 to Caltech, and NSF Grants No. PHY0354631 and PHY0312072 to Cornell University. This research used the resources of the Center for Computation and Technology at Louisiana State University, which is supported by funding from the Louisiana legislature's Information Technology Initiative. We thank Gioel Calabrese, Rob Myers, Jorge Pullin and Oscar Reula for several discussions related to the applications presented in this work.

References

1. B. Gustaffson, H. Kreiss, J. Oliger: *Time Dependent Problems and Difference Methods* (Wiley, New York 1995) 224
2. B. Strand: Journal of Computational Physics **110**, 47 (1994) 224, 225
3. P. Olsson: Summation by parts, projections and stability. I. Mathematics of Computation **64**, 1035 (1995); Supplement to Summation by parts, projections and stability. I. Mathematics of Computation **64**, S23 (1995); Summation by parts, projections and stability. II. Mathematics of Computation **64**, 1473 (1995) 224, 226
4. G. Calabrese, L. Lehner, D. Neilsen, J. Pullin, O. Reula, O. Sarbach, M. Tiglio: Class. Quant. Grav. **20**, L245 (2003) 224, 225, 226

5. G. Calabrese, L. Lehner, O. Reula, O. Sarbach, M. Tiglio: gr-qc/0308007 224
6. L. Lehner, D. Neilsen, O. Reula, M. Tiglio: gr-qc/0406116 224, 242
7. G. Calabrese, J. Pullin, O. Sarbach, M. Tiglio: Phys. Rev. D **66**, 041501 (2002) 224
8. H.-O. Kreiss, L. Wu: Appl. Numr. Math. **12**, 213 (1993) 226
9. D. Levy and E. Tadmor: SIAM Journal on Num. Anal. **40**, 40 (1998) 226, 237
10. J.M. Stewart: Class. Quantum Grav. **15**, 2865 (1998) 227
11. H. Friedrich, G. Nagy: Comm. Math. Phys. **201**, 619 (1999) 227
12. M.S. Iriondo, O.A. Reula: Phys. Rev. D **65**, 044024 (2002) 227
13. B. Szilagyi, B. Schmidt, J. Winicour: Phys. Rev. D **65**, 064015 (2002) 227
14. J.M. Bardeen, L.T. Buchman: Phys. Rev. D **65**, 064037 (2002) 227
15. G. Calabrese, L. Lehner, M. Tiglio: Phys. Rev. D **65**, 104031 (2002) 227
16. B. Szilagyi, J. Winicour: Phys. Rev. D **68**, 041501 (2003) 227
17. G. Calabrese, J. Pullin, O. Reula, O. Sarbach, M. Tiglio: Comm. Math. Phys. **240**, 377 (2003) 227
18. G. Calabrese, O. Sarbach: J. Math. Phys. **44**, 3888 (2003) 227
19. S. Frittelli, R. Gomez: Class. Quantum Grav. **20**, 2379 (2003); Phys. Rev. D **68** 044014 (2003); Phys. Rev. D **69**, 124020 (2004) gr-qc/0404070 227
20. C. Gundlach, J.M. Martín-García: Phys. Rev. D **70**, 044031 (2004); 044032 (2004) 227
21. L. Lindblom, M.A. Scheel, L.E. Kidder, H.P. Pfeiffer, D. Shoemaker, S.A. Teukolsky: Phys. Rev. D **69**, 124025 (2004) 227
22. O. Reula, O. Sarbach: gr-qc/0409027 227
23. L. Kidder, M. Scheel, S. Teukolsky: Phys. Rev. D **64**, 064017 (2001) 227
24. M. Tiglio: gr-qc/0304062 227, 229
25. E. Witten: Nucl. Phys. B **195**, 481 (1982) 230, 233, 237
26. D. Brill, H. Pfister: Phys. Lett. B **228**, 359 (1989) 230
27. D. Brill, G.T. Horowitz: Phys. Lett. B **262**, 437 (1991) 230, 231
28. S. Corley, T. Jacobson: Phys. Rev. D **49**, 6261 (1994) 232
29. O. Sarbach, L. Lehner: Phys. Rev. D **69**, 021901 (2004) 232, 237
30. R. Wald: Annals of Phys. **83**, 548 (1974) 239
31. M. Choptuik, L. Lehner, I. Olabarrieta et al: Phys. Rev. D **68**, 044001 (2003) 239
32. O. Sarbach, L. Lehner: hep-th/0407265 241, 248
33. M. Tiglio, L. Lehner, D. Neilsen: gr-qc/0312001 241, 243
34. O. Sarbach, M. Tiglio: Phys. Rev. D **66**, 064023 (2002) 241
35. K. Martel, E. Poisson: Am. J. Phys. **69**, 476 (2001) 242
36. J. Thornburg: gr-qc/0306056 246

Some Mathematical Problems in Numerical Relativity

Maria Babiuc[1], Béla Szilágyi[1], Jeffrey Winicour[1,2]

[1] Department of Physics and Astronomy, University of Pittsburgh, Pittsburgh, PA 15260, USA
maria@einstein.phyast.pitt.edu
bela@einstein.phyast.pitt.edu
[2] Max-Planck-Institut für Gravitationsphysik, Am Mühlenberg 1, 14476 Golm, Germany
jeff@einstein.phyast.pitt.edu

Abstract. The main goal of numerical relativity is the long time simulation of highly nonlinear spacetimes that cannot be treated by perturbation theory. This involves analytic, computational and physical issues. At present, the major impasses to achieving global simulations of physical usefulness are of an analytic/computational nature. We present here some examples of how analytic insight can lend useful guidance for the improvement of numerical approaches.

1 Introduction

The main goal of numerical relativity is the long time simulation of highly nonlinear spacetimes that cannot be treated by perturbation theory. There are three elements to achieving this:

(1) Analytic issues, such as well-posedness, constraints, boundary conditions, linear stability, gauge conditions and singularity avoidance.
(2) Computational issues, such as evolution and boundary algorithms, numerical stability, consistency, spacetime discretization and numerical dissipation.
(3) Physical issues, such as simulation of the desired global spacetime, extraction of the radiation from an isolated system, the proper choice of initial data, long timescale evolutions tracking many orbits of an inspiraling binary.

The correct treatment of the physical issues introduces severe global problems. For instance, long term simulations are needed to flush out the spurious gravitational radiation contained in the initial data for the gravitational field of a binary black hole so that a physically relevant waveform can be extracted. Furthermore, extraction of the waveform requires a compactified grid extending to null infinity, or some alternative approximation based upon an outer boundary in the radiation zone.

M. Babiuc et al.: *Some Mathematical Problems in Numerical Relativity*, Lect. Notes Phys. **692**, 251–274 (2006)
www.springerlink.com

At present, the major impasses to achieving such global simulations are of an analytical/computational nature. We present here some examples of how analytic insight can lend useful guidance for the improvement of numerical approaches.

2 Waves

The prime physical objective is to compute the gravitational waves emanating from a compact source. We begin by introducing some underlying mathematical and computational problems in terms of two examples of scalar wave propagation. Both of these examples are chosen because they have a direct analogue in general relativity and illustrate computational problems that arise because of exponentially growing modes in an analytic problem which is well-posed.

2.1 Unbounded Exponential Growth

First, consider a nonlinear wave propagating in Minkowski space according to

$$\eta^{\alpha\beta}\partial_\alpha\partial_\beta\Phi - \frac{1}{\Phi}\eta^{\alpha\beta}(\partial_\alpha\Phi)(\partial_\beta\Phi) = 0 = \Phi\eta^{\alpha\beta}\partial_\alpha\partial_\beta\log\Phi\ . \tag{1}$$

Although this nonlinear equation arises from a linear wave equation for $\log\Phi$, it is a remarkably accurate model for some of the problems that occur in numerical relativity. In order to simplify the problem, we first impose periodic boundary conditions so that the evolution takes place in a 3-torus T^3, i.e. on a boundary free manifold. For $\Phi > 0$ the Cauchy problem for this system is well-posed. Furthermore, the linear superposition $\log\Phi_1 + \log\Phi_2$ of solutions to the linear wave equation correspond to the solution $\Phi_1\Phi_2$ of the nonlinear problem.

A nonsingular solution of this system is the wave

$$\Phi = 1 + F(t - z)\ , \tag{2}$$

where $F > -1$. Suppose we try to simulate this solution numerically. If numerical error excites an exponentially growing mode of this system then noise in this mode will eventually dominate the wave being simulated. For the linear wave equation there are no such exponential modes. But this nonlinear system admits the solutions

$$\Phi_\lambda = e^{\lambda t}(1 + F(t - z))\ , \tag{3}$$

for arbitrary λ.

Thus, although we have a well-posed initial value problem whose principle part is the Minkowski wave operator, the simulation of a simple traveling

wave is complicated by the existence of neighboring solutions which grow exponentially in time. Numerical error will excite these exponentially growing modes and eventually dominate the traveling wave we are attempting to simulate. You might ask: Why not choose $\log \Phi$ as the evolution variable? This would clearly solve all numerical problems. As we have already said, we have chosen this example because it arises in numerical relativity where there is no analogous way to take the logarithm of the metric. But there are indirect ways to model the derivative of a logarithm, analogous to grouping terms according to $\Phi^{-1}\partial\Phi$ and rewriting the nonlinear wave equation (1) as

$$\eta^{\alpha\beta}\partial_\alpha\left(\frac{1}{\Phi}\partial_\beta\Phi\right) = 0\,. \tag{4}$$

This indeed works for the scalar field, as illustrated in Fig. 2. (See the discussion of finite difference methods in Sect. 3.) We will come back to the gravitational version of this problem but first we give another example which illustrates a similar complication with the *initial-boundary* value problem.

We base this example on the initial-boundary value problem (IBVP) for this nonlinear scalar field in the region $-L \le z \le L$ obtained by opening up the 3-torus to $T^2 \times R$. We consider the simulation of a traveling wave packet $\Phi = 1 + F(t-z) > 0$ with the Neumann boundary condition $\partial_z\Phi|_{z=\pm L} = \partial_z F|_{z=\pm L}$. The wave packet gets the correct Neumann boundary data for it to enter the boundary at $z = -L$, propagate across the grid and exit the boundary at $z = +L$.

In the process, numerical noise will be generated. There are solutions of the system of the form

$$\Phi_\epsilon = \epsilon^2 e^{t/\epsilon}\Big(1 + f(t-z)\Big)\,, \tag{5}$$

for arbitrary $\epsilon > 0$. Normally, if a scalar field admitted such solutions we could infer that the corresponding Cauchy problem was ill-posed by arguing (following Hadamard) that the solution $\Phi_0 = 0$ has vanishing Cauchy data and that the neighboring solutions Φ_ϵ have unbounded size for any $t > 0$. However, the above Cauchy problem is well-posed because the initial data must satisfy $\Phi > 0$.

After the wave packet has crossed the grid, the remnant numerical noise gets homogeneous Neumann data $\partial_z\Phi|_{z=\pm L} = 0$. Thus it is reflected off the boundaries and trapped in the simulation domain where it can grow exponentially. The noise is generated while the wave packet is traveling across the grid. The short wavelength modes can be controlled by introducing numerical dissipation. But the long wavelength modes cannot be damped without the danger of damping the signal. Just as in the case of periodic boundary conditions, numerical error can excite exponential modes that destroy the accuracy of a simulation in the case of Neumann boundary conditions.

You might ask: Why use Neumann boundary conditions? The Sommerfeld boundary condition $(\partial_t \pm \partial_z)\Phi|_{z=\pm L} = 0$ does not admit such modes and

moreover it propagates numerical noise off the grid. The answer to that question has to wait until we have discussed the constraint equations of general relativity.

The problem with using Neumann boundary conditions is not of analytic origin. The problem is of a numerical nature. Whereas the signal gets the correct inhomogeneous boundary data to propagate it off the evolution domain, the noise gets the left-over homogeneous data and gets trapped in exponentially growing modes.

The lesson here is that it is preferable to use Sommerfeld type boundary conditions, not for physical or mathematical advantage but for numerical advantage. A Sommerfeld boundary condition doesn't solve all the problems. Even though a homogeneous Sommerfeld condition carries energy out of the evolution domain, it does not in general give the physically correct outer boundary condition for an isolated system. For a nonlinear system such as general relativity, one would need an inhomogeneous Sommerfeld condition whose boundary data could only be determined by matching to an exterior solution. But numerically a Sommerfeld condition has the great advantage of allowing the noise to escape through the boundary. Unfortunately, in present numerical relativity codes, Sommerfeld boundary conditions are inconsistent with the constraints, which we will discuss later.

2.2 Moving Boundaries

Another mechanism by which a reflecting boundary condition can introduce exponential modes is the repetitive blue shifting off moving boundaries. This can even happen for a linear wave propagating between two plane boundaries in Minkowski space. The boundaries can effectively play ping-pong with a wave packet by arranging the boundary motion to be always toward the packet during reflection.

Let $\hat{x}^\alpha = (\hat{t}, \hat{x}, \hat{y}, \hat{z})$ be inertia coordinates, with the reflecting boundaries in the $(\hat{x}, \hat{y})$ plane. Under reflection, functional dependence of a wave packet traveling in the positive $\hat{z}$-direction changes according to

$$\Phi(\hat{t} - \hat{z}) \to \Phi(e^{2\alpha}(\hat{t} + \hat{z})) \,, \tag{6}$$

where the speed of the reflecting wall is

$$v = \tanh \alpha \,. \tag{7}$$

After many reflections the energy in the wave grows by a factor $e^{4\alpha T}$, where T is measured in units of the crossing-time between reflections.

It is instructive to reinterpret this experiment from a numerical relativity viewpoint where the spatial coordinates of the boundaries have fixed grid values. For that purpose, we consider the well-posedness of the IBVP for the linear wave equation

$$g^{\alpha\beta}\partial_\alpha\partial_\beta\Phi = 0 \tag{8}$$

in a general background spacetime with non-constant metric $g_{\alpha\beta}$. Again let the evolution domain be the region $-L \le z \le L$.

Most of the mathematical literature on well-posedness of the IBVP is based upon symmetric hyperbolic systems in first derivative form [1, 2, 3, 4]. We achieve this for the wave equation by introducing auxiliary variables $\mathbf{u} = (\Phi, \partial_\alpha\Phi)$. Standard results then imply a well posed IBVP for a homogeneous boundary condition of the matrix form $\mathbf{M}\mathbf{u} = 0$ provided that

- the resulting energy flux normal to the boundary has the *dissipative* property
$$\mathcal{F}^n(\mathbf{u}) \ge 0\,, \tag{9}$$
- $\mathbf{M}$ has *maximal* rank consistent with this dissipative property, and
- $\mathbf{M}$ is independent of $\mathbf{u}$.

In the present case, the energy flux is determined by the energy momentum tensor for the scalar field. The flux normal to the boundary at $z = +L$ is

$$\mathcal{F}^n = -n^\alpha(\partial_t\Phi)\partial_\alpha\Phi \tag{10}$$

where $n^\alpha = g^{z\alpha}/\sqrt{g^{zz}}$ is the unit normal to the boundary.

The dissipative condition can be satisfied in a variety of ways. The choice

$$\partial_t\Phi = 0 \tag{11}$$

leads to a homogeneous Dirichlet boundary condition, and the choice

$$n^\alpha\partial_\alpha\Phi = 0 \tag{12}$$

leads to a homogeneous Neumann boundary condition. Homogeneous Dirichlet and Neumann boundary conditions are limiting cases for which $\mathcal{F}^n = 0$, i.e. there is no energy flux across the boundary and signals are reflected. Between these limiting cases, there is a range of homogeneous boundary conditions with $\mathcal{F}^n > 0$ of the form $n^\alpha\partial_\alpha\Phi + P\partial_t\Phi = 0$, where $P > 0$. Of particular interest is the Sommerfeld-like case where the derivative lies in an outgoing characteristic direction.

The IBVP for the scalar wave equation is well-posed for any of these *maximally dissipative* boundary conditions. Furthermore, by consideration of the symmetric hyperbolic equation satisfied by $\mathbf{u} - \mathbf{q}(x^\alpha)$, where $\mathbf{q}$ has explicitly prescribed space-time dependence, the well-posedness of the IBVP with the homogeneous boundary condition $\mathbf{M}\mathbf{u} = 0$ can be extended to the inhomogeneous form $\mathbf{M}(\mathbf{u} - \mathbf{q}(t, x, y)) = 0$, with freely assigned boundary data $\mathbf{q}$. By using such boundary data, a Neumann or Dirichlet boundary condition can be used to model a wave which is completely transmitted across the boundary with no reflected component, at least at the analytic level.

Note the important feature that the free boundary data for the scalar field consist of one function of three variables in contrast to two functions for free

Cauchy data. As we shall see, this is the major complication in formulating a constraint preserving boundary condition for a well-posed IBVP in general relativity.

Now consider the simulation of a linear scalar wave in the flat background metric which results from the transformation $\hat{t} = t$, $\hat{x} = x$, $\hat{y} = y$, $\hat{z} = z + A \sin \omega t$ from inertial coordinates $\hat{x}^\alpha$. In these x^α coordinates, the boundaries at $z = \pm L$ are oscillating relative to the inertial frame, as indicated by the "shift" $g^{zt} = -A\omega \cos \omega t$. For our simulation, we prescribe initial data $\Phi_0 = \partial_t \Phi_0 = 0$ and either the appropriate Neumann or Dirichlet boundary data $q_{-L}(t)$ and $q_{+L}(t)$ for a wave packet which enters the boundary at $z = -L$, travels across the domain and leaves the boundary at $z = +L$. A second order accurate code would simulate this signal with $O(\Delta^2)$ error in the grid displacement Δ. Thus $\Phi = O(\Delta^2)$ after the packet has traversed the domain. However, this remnant error gets homogeneous boundary data. Although, as just discussed, the normal energy flux associated with homogeneous Dirichlet or Neumann data vanishes in the rest frame of the boundary, in the $\hat{x}^\alpha$ inertial frame the boundary is moving and the noise can be repeatedly blue shifted, resulting in an exponential increase of energy.

One way to eliminate this problem would be to deal with coordinate systems in which the shift vanishes, at least at the boundary. However, this would rule out many promising strategies for dealing with binary black holes, e.g. the use of co-rotating coordinates or of generalized Kerr-Schild coordinates. But especially in a nonlinear problem such as general relativity, the excitation of exponential modes can rapidly destroy code performance.

3 General Relativity: Harmonic Evolution

The previous examples of scalar waves show that even if the underlying analytic problem is well-posed and even if the numerical simulation converges to the analytic solution, the existence of exponentially growing modes in the analytic system can effectively invalidate long term code performance. In general relativity, coordinate freedom is a further complication which can introduce rapidly growing modes that are an artifact of gauge pathologies. In order to illustrate computational problems that are not a trivial consequence of gauge, we consider the harmonic formulation of Einstein's equations. Although no coordinates can guarantee complete avoidance of gauge problems, harmonic coordinates have several advantages for investigating the interface between numerical and analytic problems in general relativity:

- Small number of variables
- Small number of constraints (4 harmonic conditions)
- Einstein's equations reduce to quasilinear wave equations
- Well-posed Cauchy problem [5]
- Symmetric hyperbolic formulation [6]

– Global asymptotically flat solutions for weak Cauchy data [7]
– Well-posed homogeneous IBVP [8]

A numerical code for evolving Einstein's equations, the Abigel code [8], has been based upon a generalized version of harmonic coordinates satisfying the curved space wave equation

$$H^\alpha := \sqrt{-g}\Box_g x^\alpha = \partial_\mu(\sqrt{-g}g^{\mu\nu}\partial_\nu x^\alpha) = \tilde{H}^\alpha(x^\beta, g_{\rho\sigma}) \,, \tag{13}$$

where $\tilde{H}^\alpha$ are harmonic source terms. These harmonic source terms do not affect any of the analytic results regarding well-posedness but, in principle, they allow any spacetime to be simulated in some generalized harmonic coordinate system. The harmonic reduced evolution equations are written in terms of the metric density $\gamma^{\mu\nu} = \sqrt{-g}g^{\mu\nu}$ whose ten components obey quasilinear wave equations

$$\gamma^{\alpha\beta}\partial_\alpha\partial_\beta\gamma^{\mu\nu} = S^{\mu\nu}, \tag{14}$$

where $S^{\mu\nu}$ contains nonlinear first-derivative terms that do not enter the principal part. The harmonic conditions $C^\mu := H^\mu - \tilde{H}^\mu = 0$ are the constraints on this evolution system which are sufficient to ensure that Einstein's equations are satisfied. Except where noted, we set $\tilde{H}^\mu = 0$ to simplify the discussion but all results generalize to include nonvanishing gauge source terms. For details concerning the formulation and its implementation see [8].

By virtue of the evolution equations, the harmonic constraints satisfy a homogeneous wave equation of the form

$$\gamma^{\alpha\beta}\partial_\alpha\partial_\beta C^\mu + A^{\mu\alpha}_\beta\partial_\alpha C^\beta + B^\mu_\beta C^\beta = 0 \,. \tag{15}$$

Thus, in the domain of dependence of the Cauchy problem, the solution $C^\mu = 0$ is implied by standard uniqueness theorems provided the system is initialized correctly.

A well-posed evolution system is a necessary but not sufficient ingredient for building a reliable evolution code. Code performance can be best tested by simulating an exact solution and measuring an error norm. The error should converge to zero in the continuum limit as the grid spacing Δ shrinks to zero. In testing evolution codes it is desirable to first eliminate effects of boundary conditions by imposing periodicity in space, which is equivalent to carrying out the simulation on a 3-torus without boundary. A suite of toroidal testbeds for numerical relativity has been developed as part of the Apples with Apples project [9, 10]. The convergence and stability of several codes [9, 11, 12], including the Abigel code [9], has been demonstrated using this test suite.

One testbed is the Apples with Apples periodic gauge wave with metric

$$ds^2 = \Phi(-dt^2 + dz^2) + dx^2 + dy^2, \tag{16}$$

where

$$\Phi = 1 + A \sin\left(\frac{2\pi(t-z)}{2L}\right) . \tag{17}$$

It is obtained from the Minkowski metric $ds^2 = -d\hat{t}^2 + d\hat{x}^2 + d\hat{y}^2 + d\hat{z}^2$ by the harmonic coordinate transformation

$$\begin{aligned} \hat{t} &= t - \tfrac{AL}{\pi} \cos\left(\frac{2\pi(t-z)}{2L}\right), \\ \hat{z} &= z + \tfrac{AL}{\pi} \cos\left(\frac{2\pi(t-z)}{2L}\right), \\ \hat{x} &= x\,, \\ \hat{y} &= y\,. \end{aligned} \tag{18}$$

Figure 1 shows some snap shots of a gauge wave simulation carried out with an early version of the Abigel code. The time dependence of the error shows exponential growth of the form

$$\mathcal{E} \sim \Delta^2 \psi(t,z) e^{\lambda t}. \tag{19}$$

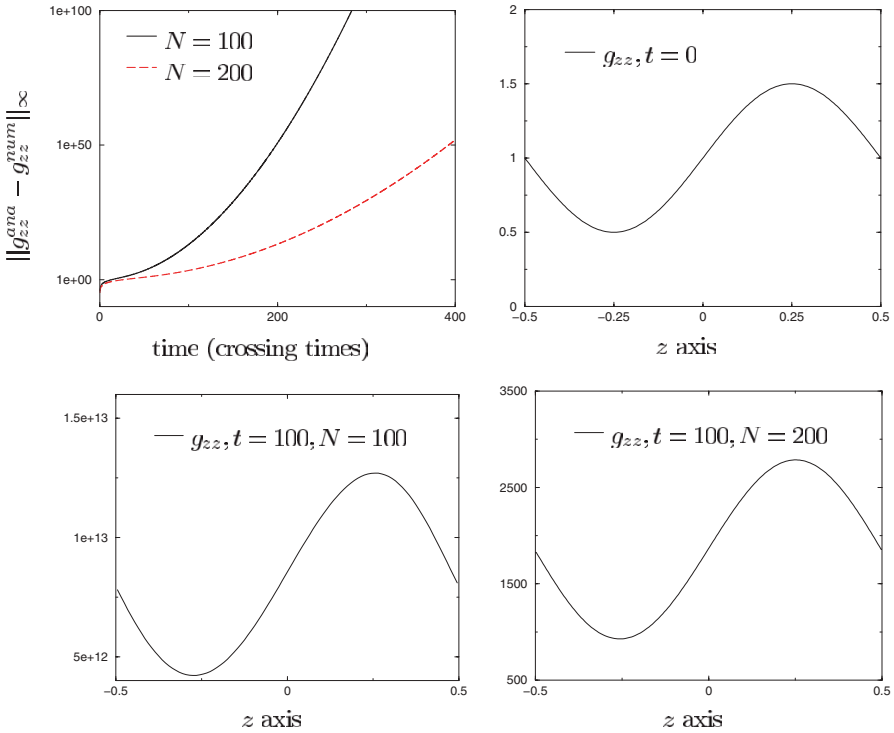

Fig. 1. Snap shots of a gauge wave simulation carried out with an early version of the Abigel code. The code is stable and the error converges to zero at second order in grid spacing Δ but after a few crossing times the error is too large to make the results useful

Here ψ is a well-behaved function which is almost identical in shape to the signal Φ. As a result, the error cannot be dissipated by standard techniques for dealing with short wavelength noise. The exponential growth originates in exact analogy with our example for the nonlinear scalar wave equation (1). In fact, the metric (16) is a flat solution of the Einstein equations in harmonic coordinates for any $\Phi(t, z)$ which satisfies the nonlinear wave equation (1). The general solution is $\Phi = e^{f(t-z)+g(t+z)}$. In particular, there are exponentially growing harmonic gauge waves

$$ds^2 = \Phi_\lambda(-dt^2 + dz^2) + dx^2 + dy^2, \tag{20}$$

where

$$\Phi_\lambda = e^{\lambda t}\left(1 + A\sin\left(\frac{2\pi(t-z)}{2L}\right)\right), \tag{21}$$

which lie arbitrarily close to the gauge wave being simulated. Thus numerical error inevitably excites exponential modes which eventually dominate the simulation of the gauge wave. The practicality of code performance depends on the timescale of this exponential growth.

Although the Abigel code is stable, convergent and based upon a well-posed initial value problem, like any other code it is subject to the excitation of exponential modes in the underlying analytic system. Certain numerical techniques greatly improve its accuracy in simulating the gauge wave:

- Expressing the equations into *flux conservative form*, an idea from computational fluid dynamics which was introduced into general relativity by the Palma group [16].
- *Summation by parts*, introduced into general relativity by the LSU group [12], which at the level of linearized equations leads to energy estimates for the semi-discrete system of ODE's in time which arise from spatial discretization.
- *Nonlinear multipole conservation*, which suppresses the excitation of long wavelength exponential modes by grouping the troublesome nonlinear terms in a way that enforces global semi-discrete conservation laws (or approximate conservation laws).

The semi-discrete multipole technique (being introduced here) provides an excellent example of how analytic insight into the source of a numerical problem can be used to design a remedy. As we will show, various combinations of these three techniques lead to dramatic improvement in gauge wave simulations. Other numerical methods based upon enforcing or damping the constraints are not crucial for the gauge wave problem but can be important for simulations of curved spacetimes.

An essential ingredient in any code is the method used to approximate derivatives. The Abigel code treats the quasilinear wave equations (14) as first differential order in time and second order in space. This allows use of

explicit finite difference methods to deal with the mixed space-time derivatives introduced by the "shift" term in the wave operator while avoiding the artificial constraints that would be introduced by full reduction to a first order system. On a grid with spacing Δ, the natural finite difference representation for the the first and second spatial derivatives are the centered approximations

$$\partial_z F(z) \to DF(z) = \frac{F(z+\Delta) - F(z-\Delta)}{2\Delta} \tag{22}$$

and

$$\partial_z^2 F(z) \to D^{(2)} F(z) = \frac{F(z+\Delta) - 2F(z) + F(z-\Delta)}{\Delta^2} \,. \tag{23}$$

These formulae were used in the simulation labeled TIGHT in Fig. 3. Although the code was tested to be stable and convergent with second order accuracy, the excitation of the exponentially growing mode of the analytic problem limits accurate simulations to about 10 crossing times on a reasonably sized grid.

In order to exaggerate nonlinear effects, the simulations shown in Fig. 3 were carried out for a highly nonlinear gauge wave with amplitude $A = .5$, on a scale where the metric is singular for $A = 1$. (The standard Apples with Apples tests specify amplitudes of $A = .01$ and $A = .1$.) Problems with exponential modes do not appear for small amplitudes simulations in the linear domain. One contributing factor to the exponential growth is that the tight 3-point stencil (23) for the second derivative does not lead to an exact finite difference representation of the integration by parts rule necessary to establish energy conservation, which is the main idea behind the summation by parts (SBP) method. But this is only part of the story since standard SBP techniques only apply to linear systems.

It is instructive to examine how these ideas extend to the second derivative form of the nonlinear wave equation (1) which underlies the gauge wave problem. This will illustrate in a simple way how flux conservative equations, SBP and multipole conservation can combine to suppress excitation modes in the analytic problem. The model scalar problem is effective in isolating the difficulties underlying a full general relativistic code, in addition to allowing efficient computational experimentation.

We carry out the analysis for waves traveling with periodic boundary conditions in one spatial dimension. The extension to three dimensions is straightforward but notationally more complicated. The theory regarding well-posedness of hyperbolic systems is based upon the principal part of the equations. For that reason, we first consider the linear wave equation

$$\partial_\alpha \partial^\alpha \Phi = -\partial_t^2 \Phi + \partial_z^2 \Phi = 0 \,. \tag{24}$$

The energy associated with this wave can be related to the conserved integral

$$\mathcal{I} = [\Phi_1, \Phi_2] = \oint (\Phi_1 \partial^\mu \Phi_2 - \Phi_2 \partial^\mu \Phi_1) dV_\mu \tag{25}$$

by choosing $\Phi_1 = \Phi$ and $\Phi_2 = \partial_t \Phi$. For the case of periodic boundary conditions on the interval $0 \leq z \leq L$,

$$\mathcal{I} = \int_0^L \Big((\partial_t \Phi)^2 - \Phi \partial_t^2 \Phi \Big) dz \; . \tag{26}$$

The integration by parts by parts rule

$$\int_0^L \Big(- \partial_z (\Phi \partial_z \Phi) + (\partial_z \Phi)^2 + \Phi \partial_z^2 \Phi \Big) dz = 0 \; , \tag{27}$$

applied to a periodic interval, then supplies the key step in using the wave equation to relate $\mathcal{I}$ to the positive definite energy

$$\mathcal{I} = \mathcal{E} = \int_0^L \Big((\partial_t \Phi)^2 + (\partial_z \Phi)^2 \Big) dz \; . \tag{28}$$

In order to obtain a discrete version of the integration by parts identity (27), we introduce a uniform grid z_i , $0 \leq i \leq N$, with spacing Δ and represent

$$\int_0^L F dz \to \Delta \sum_0^N f_{i+1/2} \tag{29}$$

where

$$f_{i+1/2} = \frac{F(z_i) + F(z_{i+1})}{2} \; . \tag{30}$$

In addition we represent derivatives at the midpoints by the centered approximation

$$\partial_z F(z_i + \Delta/2) \to f'_{i+1/2} = \frac{F(z_{i+1}) - F(z_i)}{\Delta} \tag{31}$$

so that periodic boundary conditions imply

$$\int_0^L \partial_z F dz \to \Delta \sum_0^N f'_{i+1/2} = F|_0^L = 0 \; . \tag{32}$$

This ensures the semi-discrete monopole conservation law

$$\partial_t^2 \oint \Phi dz \to 0 \; , \tag{33}$$

which results from the flux conservative form of (24). Equation (33) controls growth of the spatial average of Φ but not of its non-constant spatial Fourier components which measure its gradient.

Energy estimates control the growth of the gradient of Φ. With the above definitions, it is straightforward to check that

$$\partial_z (FG) - F \partial_z G - G \partial_z F \to 0 \; . \tag{34}$$

As a result, the semi-discrete version of the integral identity (27) is satisfied if the second derivative term is represented as a product of first derivatives. For the linear wave equation this results in the semi-discrete conservation law

$$\partial_t \mathcal{E} \to 0 \,. \tag{35}$$

In order to implement SBP in a code such as the Abigel code, which is second differential order in space with the fields represented by their values on grid points, and not on mid-points, the above results can be applied by treating the mid-points for even numbered grid points as the odd-numbered grid points, and vice versa. This results in the widened finite difference representation for the second derivative

$$\partial_z^2 F(z) \to D^2 F(z) = DDF(z) = \frac{F(z+2\Delta) - 2F(z) + F(z-2\Delta)}{4\Delta^2} \,, \tag{36}$$

as opposed to the tight stencil (23). Fig. 2 shows the remarkable improvement in long term accuracy obtained in the simulation of a non-linear wave satisfying (1). (Numerical dissipation has been used to damp short wavelength instabilities triggered by the loose coupling between even and odd grid points.) The curves labeled TIGHT are obtained using the standard stencil (23). They show exponential growth on a scale of ≈ 10 grid crossing times. The curves labeled SBP are obtained using the stencil (36) consistent with SBP. This change of stencil suppresses growth of long wavelength exponential modes so that accurate simulations of ≈ 1000 crossing times are possible.

Summation by parts only has general applicability to linear equations, although the technique extends in an approximate sense to the nonlinear domain. Other approaches can also be successful for nonlinear problems, especially if the troublesome nonlinear terms can be identified. In the case of the nonlinear equation (1), these terms can be incorporated in the principal part by reformulating the equations in the flux conservative form

$$\partial_\alpha \left(\eta^{\alpha\beta} \frac{1}{\Phi} \partial_\beta \Phi \right) = 0 \,, \tag{37}$$

with the subsequent reduction to the first order in time system

$$\partial_t \Phi = \Phi Q \,, \tag{38}$$

$$\partial_t Q = \partial_z \left(\frac{1}{\Phi} \partial_z \Phi \right) \,. \tag{39}$$

Many choices of spatial discretization of this flux conservative system lead to an *exact* semi-discrete version of the monopole conservation law

$$\partial_t \int_0^L Q = 0 \,. \tag{40}$$

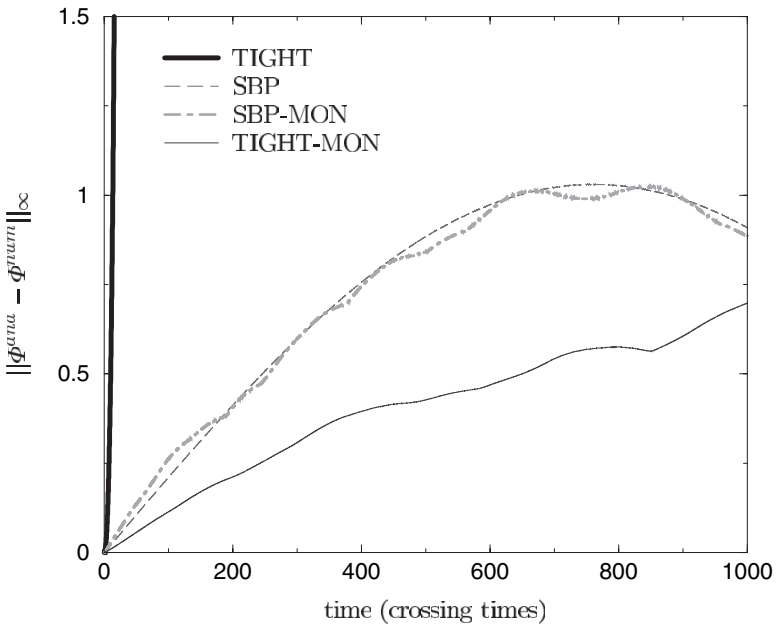

Fig. 2. A comparison of the various evolution algorithms used to evolve the nonlinear wave equation (1). The tests are based on the sine wave solution (17), with amplitude $A = .5$, simulated on a grid of $N = 100$ points with a time-step of $\Delta t = \Delta z/4$. The graph shows the ℓ_∞ norm of the error

We consider the two choices

$$\partial_z(\frac{1}{\Phi}\partial_z\Phi)|_i \to \frac{1}{2\Delta}\left(\frac{1}{\Phi}\Phi'\right)_{i+1} - \frac{1}{2\Delta}\left(\frac{1}{\Phi}\Phi'\right)_{i-1} \tag{41}$$

and

$$\partial_z(\frac{1}{\Phi}\partial_z\Phi)|_i \to \frac{1}{\Delta}\left(\frac{1}{\Phi}\Phi'\right)_{i+1/2} - \frac{1}{\Delta}\left(\frac{1}{\Phi}\Phi'\right)_{i-1/2}. \tag{42}$$

As a result of either of these discretizations, the initial data determine the conserved value of the monopole moment $\int_0^L Q dz$ and the excitation of the exponential modulation (3) is thereby frozen out of the numerical evolution. In this way a tight 3-point stencil (42) can be used, as opposed to the wide 5-point stencil (41) (and the concomitant numerical dissipation) required by SBP for the second order system. The curve labeled TIGHT-MON in Fig. 2 shows how long term accuracy is dramatically enhanced by this technique, without use of numerical dissipation. The curve labeled SBP-MON shows that, in this case, no additional improvement is gained when monopole conservation is combined with SBP.

These numerical techniques introduced for the model scalar problem were formulated in a way that could be readily taken over to the gravitational

case. Although the Einstein equations can neither be linearized by taking the "logarithm of the metric" nor written in a completely flux conservative form analogous to (37), there are various ways to group derivatives which decouple the Jacobian transformation that generates the exponential mode in the gauge wave metric (20). One example is the grouping

$$g^{\alpha\mu}\partial^{\rho}g_{\mu\beta} = (\delta^{\alpha}_{t}\delta^{t}_{\beta} + \delta^{\alpha}_{z}\delta^{z}_{\beta})\frac{1}{\Phi_{\lambda}}\partial^{\rho}\Phi_{\lambda} \ , \tag{43}$$

for which expression of the principal part of the Einstein equation in the form $\partial_{\rho}(g^{\alpha\mu}\partial^{\rho}g_{\mu\beta})$ leads to the semi-discrete conservation laws

$$\partial_{t}\int_{0}^{L} g^{\alpha\mu}\partial^{t}g_{\mu\beta}dz \rightarrow 0 \tag{44}$$

for the gauge wave. The conserved quantities are comprised of multipoles of monopole (the spatial trace) and quadrupole (the trace-free part) type.

The advantage of enforcing these conservation laws is exhibited in Fig. 3. Comparison of Figs. 2 and 3 shows that SBP and multipole conservation lead to almost identically beneficial results in simulating the gauge wave as in simulating the nonlinear scalar wave. The standard 3-point stencil (TIGHT) again excites exponentially growing error on the order of 10 crossing times but

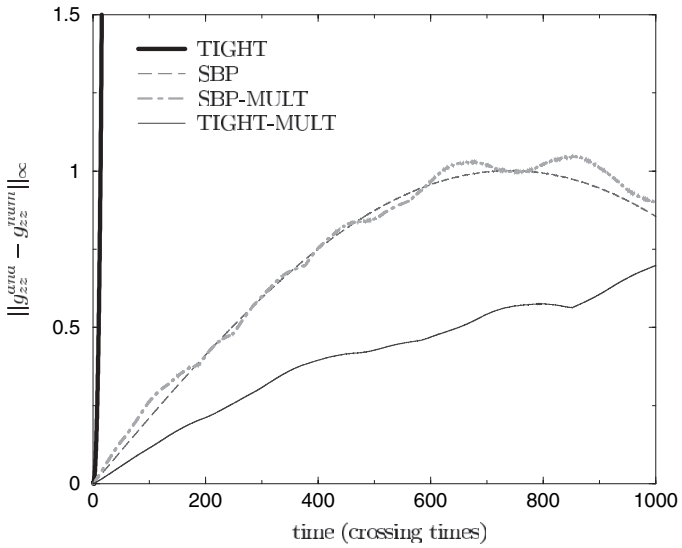

Fig. 3. A comparison of the various evolution algorithms used to evolve the harmonic Einstein equations. In these tests the code evolved flat spacetime in the gauge defined by (18), with amplitude $A = .5$. The size of the grid was $N = 100$, with a time-step of $\Delta t = \Delta z/4$. The graph shows the ℓ_{∞} error norm of the g_{zz} metric component

accurate evolutions for over 1000 crossing times are attained either with SPB or with a 3-point stencil embodying multipole conservation (TIGHT-MULT).

Whether objectionable gauge modes can be decoupled so effectively in a more general problem is an interesting question. However, all the above numerical techniques, which lead to excellent code performance for the gauge wave, are of a universal nature that can be adopted for the simulation of a general spacetime by a general code. Since most simulations contain a weak field region, such as the far field region outside a black hole, these techniques might in fact be necessary in order to avoid excitation of local versions of exponential Minkowski gauge modes. Figure 4 shows how these methods extend to the challenging simulation of the gauge wave with shift

$$ds^2 = -(1 - A\sin\alpha)dt^2 + 2A\sin\alpha dtdz + (1 + A\sin\alpha)dz^2 + dx^2 + dy^2 \ , \quad (45)$$

with $\alpha = \pi(t+z)/L$. The simulation was carried out with periodic boundary conditions and amplitude $A = .5$, so that the grid has an effective velocity of half the speed of light. Again there are exponentially growing gauge waves,

$$ds^2_\lambda = -(e^{\lambda t} - A\sin\alpha)dt^2 + 2A\sin\alpha dtdz + (e^{\lambda t} + A\sin\alpha)dz^2 + dx^2 + dy^2 \quad (46)$$

(for arbitrary λ), which satisfy these boundary conditions. These will trigger a numerical instability unless their excitation is controlled by a conservation

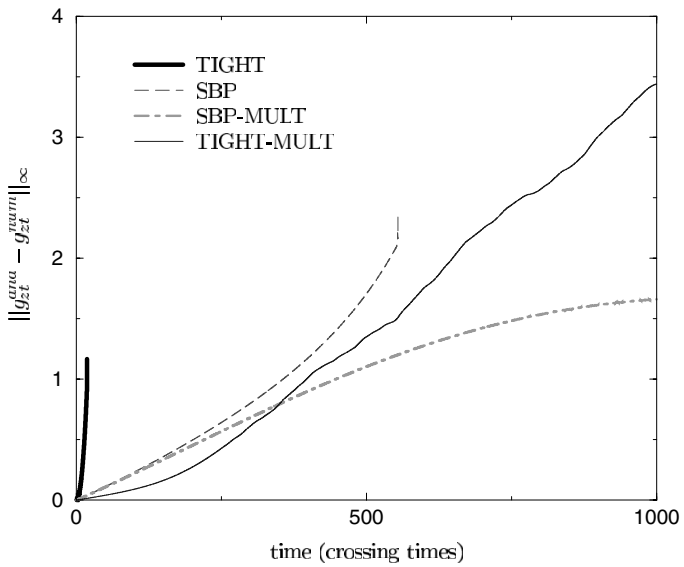

Fig. 4. A comparison of the various evolution algorithms used to evolve the harmonic Einstein equations. In these tests the code evolved the gauge wave metric with shift defined in (45), with amplitude $A = .5$. The size of the grid was $N = 100$, with a time-step of $\Delta t = \Delta z/4$

law on the semi-discrete system. Remarkable improvement in long term performance is achieved by implementing either SBP or the multipole algorithm.

These examples show what must be done, beyond having a stable, convergent code, in order to achieve accurate long term simulations. Exponential modes undoubtedly arise in a wide variety of systems with the examples presented here just the tip of the iceberg. Short wavelength modes arising from discretization error can be suppressed by numerical dissipation. The long wavelength modes exist in the analytic problem. This raises some key questions: Are there geometric clues to identify the origin of such long wavelength exponential modes? What numerical or analytic techniques can be used to suppress them?

4 The Harmonic IBVP

Given an evolution code on the 3-torus which is based upon a well-posed Cauchy problem for Einstein's equations and which is free of all numerical problems, several things can go wrong in extending the evolution to include a boundary. On the analytic side, the imposition of the boundary condition can be ill-posed or it can lead to violation of the constraints or it can introduce exponentially growing modes. On the numerical side, the finite difference implementation of the boundary condition can be unstable or inaccurate. On the physical side, the correct boundary data representing radiation (or the absence of radiation) entering the system might not be known or it might not be possible to extract the waveform of the outgoing radiation.

Here we examine the analytical and numerical issues for the harmonic IBVP. The reduced evolution system consists of the quasilinear wave equations (14). Our discussion for nonlinear scalar waves show that the IBVP for this system is well-posed for any maximally dissipative boundary conditions, e.g. Dirichlet, Sommerfeld or Neumann.

Next consider the harmonic constraints C^μ. They satisfy the homogeneous wave equation (15). Thus we can formulate a well-posed IBVP for the propagation of the constraints by imposing a maximally dissipative boundary condition. Then, given that the constraints and their time derivative are satisfied by the initial data and that the constraints have homogeneous boundary data, the uniqueness of the solution to the constraint propagation equations would imply that the constraints be satisfied in the domain of dependence of the IBVP. However, consistency between the boundary conditions for the evolution variables and the homogeneous boundary conditions for the constraints is not straightforward to arrange.

For example, consider evolution in the domain $z < 0$ with boundary at $z = 0$. In the tangential-normal $3+1$ decomposition $x^\mu = (x^a, z)$ intrinsic to the boundary, a homogeneous Dirichlet condition on the constraints takes the explicit form

$$C^z = \partial_a \gamma^{za} + \partial_z \gamma^{zz} = 0 \tag{47}$$
$$C^a = \partial_b \gamma^{ab} + \partial_z \gamma^{az} = 0\,. \tag{48}$$

A naive attempt to satisfy these conditions by boundary data on the evolution variables would involve assigning both Dirichlet (tangential) and Neumann (normal) conditions to γ^{az}, which would be an inconsistent boundary value problem.

One way to impose consistent constraint preserving boundary conditions is based upon the well-posedness of the Cauchy problem. Consider smooth Cauchy data which is locally reflection symmetric with respect to the boundary at $z = 0$. Then in some neighborhood $-L < z < L$ of the hypersurface $z = 0$ the Cauchy problem is well-posed. On $z = 0$, the local reflection symmetry implies that the evolution equations satisfy

$$\begin{aligned} \gamma^{za} &= 0 \\ \partial_z \gamma^{zz} &= 0 \\ \partial_z \gamma^{ab} &= 0 \end{aligned} \tag{49}$$

and that the constraints satisfy

$$\begin{aligned} C^z &= 0 \\ \partial_z C^a &= 0\,. \end{aligned} \tag{50}$$

It is straightforward (although algebraically complicated) to show for the harmonic IBVP that the combination of Dirichlet and Neumann boundary conditions (49) implies that the constraints satisfy the homogeneous boundary conditions (50). Thus (49) provide homogeneous constraint preserving boundary conditions for a well-posed harmonic IBVP.

Well-posedness of the IBVP extends to the case of "small" boundary data, of the form $\mathbf{M}(\mathbf{u} - \mathbf{q}(x^a)) = 0$ discussed in Sect. 3, in the sense that the prescribed data $\mathbf{q}$ is linearized off a solution with homogeneous boundary data. However, the available mathematical theorems do not guarantee well-posedness for finite boundary data. We describe below the major issues regarding constraint preserving inhomogeneous boundary conditions for the harmonic IBVP. For further details, see [8].

Part of the inhomogeneous boundary data which generalize (49) are associated with the gauge freedom corresponding to a boundary version of the "shift". By a harmonic coordinate transformation it is alway possible to set

$$\gamma^{za} = q^a(x^b)\gamma^{zz} \tag{51}$$

at the boundary, where q^a is freely prescribed data. The unit normal N^μ to the boundary then defines the normal derivative

$$\partial^n := \frac{1}{N^z} N^\mu \partial_\mu = \partial_z + q^a \partial_a \tag{52}$$

entering the Neumann boundary data, $q^{zz} = \partial^n \gamma^{zz}$ and $q^{ab} = \partial^n \gamma^{ab}$, which complete the inhomogeneous version of (49).

The boundary data $\mathbf{q} = (q^a, q^{zz}, q^{ab})$ can be freely prescribed in a well-posed IBVP for the reduced evolution system but they must be restricted to satisfy (50) in order to ensure that the constraints are satisfied. The condition $C^z = 0$ requires

$$q^{zz} = -\partial_a q^a \gamma^{zz} \ . \tag{53}$$

When the boundary shift $-q^a$ is nonzero, the second condition in (50) must be restated in the form $\partial^n C_a = 0$, because the derivative ∂_z is no longer in the normal direction to the boundary. This condition is a restriction on the data q^{ab}, which are closely related to the extrinsic curvature K^{ab} of the boundary. It requires that

$$\sqrt{-h} D_b (K^b_a - \delta^b_a K) + \sqrt{g^{zz}} K_{ab} C^b - \frac{g^{zz}}{2} C_b \partial_a q^b = 0 \ , \tag{54}$$

where h_{ab} and D_a are the metric and connection intrinsic to the boundary. This equation can be recast as a symmetric hyperbolic boundary system which determines the 6 pieces of Neumann data q^{ab} in terms of 3 free functions, the free (gauge) data q^a *and* the boundary values of γ^{zz}, γ^{ab} and $\partial_z \gamma^{za}$. *Solutions of reduced equations with this boundary data necessarily satisfy the constraints.* Unfortunately, the appearance of the quantities γ^{zz}, γ^{ab} and $\partial_z \gamma^{za}$ complicates the well-posedness of the constrained IBVP since these quantities cannot be freely specified but must be determined in the course of the evolution.

Formally, the constraint preserving boundary data have the functional dependence $\mathbf{q} = \mathbf{q}(\mathbf{u}, x^a)$, which involves evolution variables $\mathbf{u}$ whose boundary values cannot be freely prescribed. This complication has its geometric origins in the fact that the boundary data (gauge quantities and extrinsic curvature) do not include the intrinsic metric, as in the case of Cauchy data. Because of the dependence of the constraint preserving boundary data on $\mathbf{u}$, the available theorems regarding well-posedness only apply to perturbations of homogeneous data, where the background values of $\mathbf{u}$ can be explicitly determined.

These constraint preserving boundary conditions have been implemented in the Abigel code. Test simulations of the IBVP for the shifted gauge wave (45) were carried out by opening one face of the 3-torus to form a $T^2 \times [0, 1]$ manifold with boundary. Figure 5 shows the results reported for an early version of the code [8]. The graphs indicate stability and and convergence but there is also a growing error which eventually leads to a nonlinear instability. One underlying cause of this error growth is the continuous blue shifting off the moving boundaries, as discussed in Sect. 2. However, these tests were carried out before semi-discrete conservation laws were incorporated into the evolution algorithm so that a better understanding of the error must await future test runs.

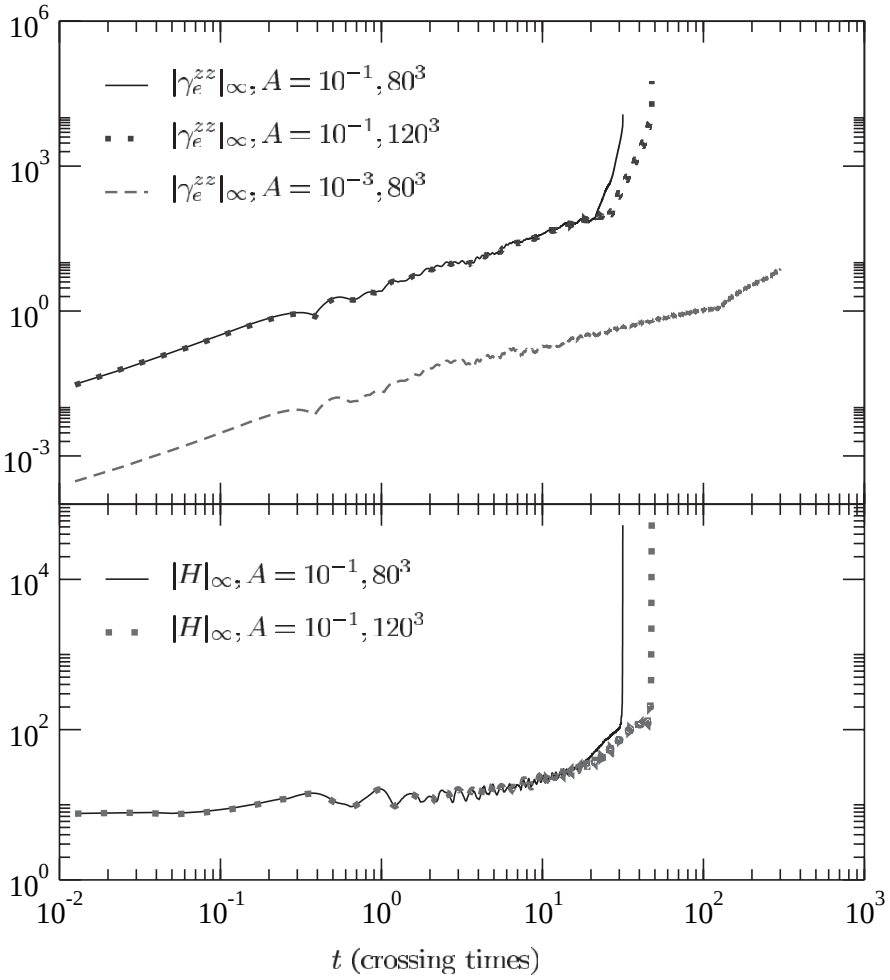

Fig. 5. The ℓ_∞ norm of the finite-difference error $\gamma_e^{zz} = \gamma_{ana}^{zz} - \gamma_{num}^{zz}$, rescaled by a factor of $1/\Delta^2$, for a gauge wave. The tests were carried out with an early version of the Abigel code before semi-discrete conservation laws were incorporated. The upper two (mostly overlapping) curves demonstrate convergence to the analytic solution for a wave with amplitude $A = 10^{-1}$ gridsizes 80^3 and 120^3. We also plot $|H|_\infty$, the ℓ_∞ norm of $\sqrt{(H^t)^2 + \delta_{ij} H^i H^j}$, to demonstrate that convergence of the harmonic constraints is enforced by the boundary conditions. The lower curve represents evolution of the same gauge wave with $A = 10^{-3}$ for 300 crossing times, with gridsize 80^3

5 Sommerfeld Alternatives

The examples presented here indicate a computational advantage in formulating boundary conditions in a manner such that numerical noise can propagate off the grid for the case of homogeneous boundary data. To date, there exists only one well-posed formulation of the IBVP for general relativity that allows this type of generalized Sommerfeld boundary condition. This is the Friedrich-Nagy formulation [13] based upon a formulation of Einstein's equations in which an orthonormal tetrad, the connection and the Weyl curvature are treated as evolution variables. The gauge freedom in the theory is adapted in a special way to the boundary so that boundary conditions need only be imposed on the curvature variables. The critical feature of the formalism is that the constraints propagate tangential to the boundary. This allows the well-posedness of the IBVP for the reduced evolution system to be extended to the fully constrained system. Unfortunately, this formulation has not yet been implemented as a numerical code, partially because of its analytic complexity and partially because it would require some infrastructure beyond that existing in most present codes.

An important issue is whether this success of the Friedrich–Nagy system in handling a Sommerfeld boundary condition is limited to formulations that include the tetrad or the curvature among the basic evolution variables. In linearized gravitational theory, there is a simple variant of the harmonic formulation that has a well-posed IBVP, admits a Sommerfeld boundary condition and has been successfully implemented computationally [15]. The nonlinear counterpart consists of the evolution system

$$\gamma^{\alpha\beta}\partial_\alpha\partial_\beta\gamma^{ij} = S^{ij}, \tag{55}$$

$$H^\alpha := \partial_t\gamma^{t\alpha} + \partial_j\gamma^{j\alpha} = \hat{H}^\alpha(x,\gamma)\ , \tag{56}$$

comprised of the wave equations (14) for the six spatial components γij and propagation equations for the time components $\gamma^{t\alpha}$. Alternatively, the propagation equations could be reformulated as

$$\partial_t H^\alpha = \partial_t \hat{H}^\alpha \tag{57}$$

in order to make the evolution system uniformly second differential order. Well-posedness of the nonlinear Cauchy problem does not follow in any direct way from standard theorems. An analysis of the principal part shows that this naive harmonic system is only weakly hyperbolic, which opens the door for lower derivative terms to produce instabilities [14].

It is instructive to investigate the performance of a code based upon this weakly hyperbolic harmonic system by using the Apples with Apples testbed. Figure 6 shows the results of the robust stability test, where a simulation in the linear regime is carried out with random (constraint violating) initial data. The results show an exponential rise in the violation of the Hamiltonian

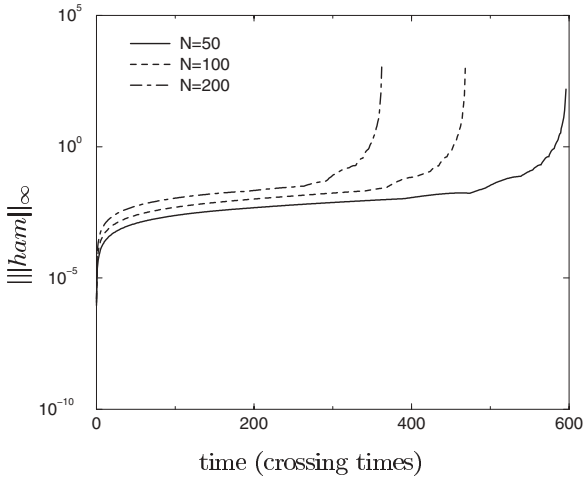

Fig. 6. The robust stability test for the weakly hyperbolic harmonic system. The ℓ_∞ norm of the Hamiltonian constraint is plotted on a linear-logarithmic scale. All specifications are in accord with the Apples With Apples test

constraint *at a rate that increases with grid resolution*, which eventually leads to a code crash. This behavior is symptomatic of weakly hyperbolic systems and presages possible problems in the nonlinear domain. The simulation of a nonlinear gauge wave with shift, shown in Fig. 7, verifies such problems. These problems do not appear for the nonlinear gauge wave without shift, as the results shown in Fig. 8 indicate convergence. Also, as illustrated in Fig. 9, with the addition of numerical dissipation, the Hamiltonian constraint no longer grows exponentially in the robust stability test, although the constraint violation still increases with grid resolution, indicating failure of the test. Similar conclusions follow from the Apples with Apples Gowdy wave tests.

These results show that a full battery of tests are necessary in order to establish reliable code performance. Otherwise, misleading information about code performance can result. As history has shown in the case of ADM evolution codes, weakly hyperbolic systems system cannot be expected to give reliable long term performance in the presence of strong fields, which makes them unsuitable for black hole simulations.

The Friedrich–Nagy system and the weakly hyperbolic harmonic system represent two extremes of a dilemma facing code development in numerical relativity. On one hand, the Friedrich–Nagy system has all the desired analytic features but its complexity poses a barrier to code development. On the other hand, the weakly hyperbolic harmonic system is simple and easily implemented as an efficient code, but well-posedness is questionable. Should you try to fix these simple systems or should you bite the bullet and develop codes based upon formulations where a well-posed nonlinear IBVP has been fully established? To date, the effort in numerical relativity has been

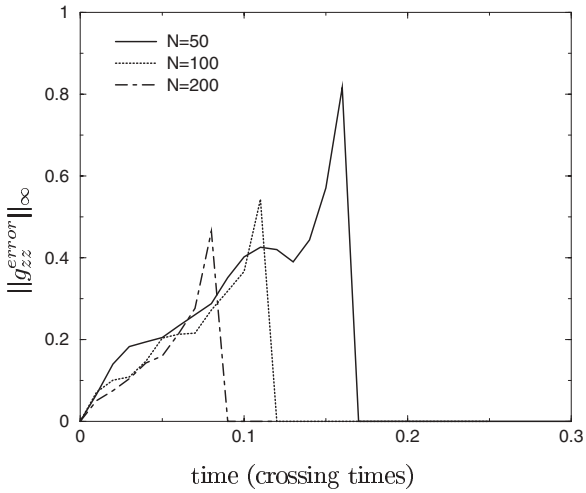

Fig. 7. The nonlinear gauge wave with shift test for the weakly hyperbolic harmonic system. The code crashes in less than a crossing time

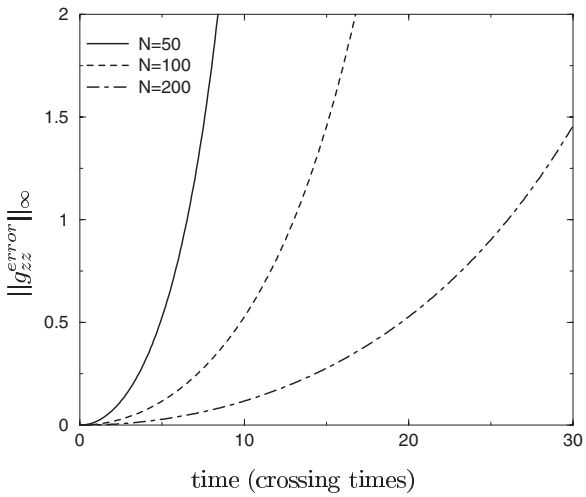

Fig. 8. The nonlinear gauge wave without shift test for the weakly hyperbolic harmonic system, run in accord with the Apples With Apples specifications. The convergence of the error is deceptive of code reliability

weighted heavily toward the simpler formulations. It is timely that some attention be given to investigating whether the Friedrich–Nagy system can be converted into a workable code. A useful starting point would be a linearized version evolving on $T^2 \times R$, where the complications of the equations and the boundary gauge would greatly simplify and would perhaps lead to a better understanding of the essential elements of the approach. Most of the effort in

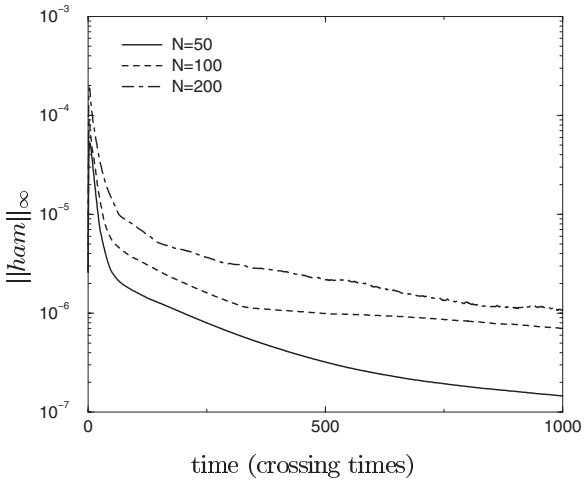

Fig. 9. The robust stability test for the weakly hyperbolic harmonic system with dissipation. As in Fig. 6, the Hamiltonian constraint is plotted on a linear-logarithmic scale. The dissipation now kills the exponential growth but the growth of constraint violation with resolution indicates that the code fails the test

the field can be expected to remain a compromise between these extremes, e.g. the strongly hyperbolic harmonic system for which a Sommerfeld boundary condition is not constraint preserving. In all such endeavors, a close working combination of analytic and numerical insight can offer valuable guidance.

Acknowledgements

Much of the material of this contribution originated from interactions with the groups carrying out the Apples with Apples standardized tests. We are particularly grateful to L. Lehner, C. Palenzuela and M. Tiglio for sharing ideas and results. These interactions demonstrate how code comparison can be very effective in tackling numerical problems. We thank B. Schmidt for several useful comments on the manuscript. The computer simulations were carried out using Cactus infrastructure. The hospitality of the Heraeus foundation at the Physikzentrum Bad Honnef during part of this work was greatly appreciated. The research was supported by National Science Foundation Grant PHY-0244673 to the University of Pittsburgh.

References

1. K.O. Friedrichs: Comm. Pure Appl. Math. **11**, 333 (1958) 255
2. P.D. Lax, R. S. Phillips: Comm. Pure Appl. Math. **13**, 427 (1960) 255

3. B. Gustafsson, H.-O. Kreiss, J. Oliger: *Time Dependent Problems and Difference Methods* (Wiley, New York 1995) 255
4. P. Secchi: Arch. Rational Mech. Anal. **134**, 155 (1996) 255
5. Y. Foures-Bruhat: Acta. Math. **88**, 141 (1955) 256
6. A.E. Fischer, J.E. Marsden: Comm. Math. Phys. **28**, 1 (1972) 256
7. H. Lindblad, I. Rodnianski: Global existence for the Einstein vacuum equations in wave coordinates, `AP/0312479` 257
8. B. Szilágyi, J. Winicour: Phys. Rev. D **68**, 041501 (2003) 257, 267, 268
9. `www.appleswithapples.org` 257
10. M. Alcubierre et al: Class. Quantum Grav. **21**, 589 (2004) 257
11. C. Bona, T. Ledvinka, C. Palenzuela, M. Zacek: A symmetry-breaking mechanism for the Z4 general-covariant evolution system. `gr-qc/0307067` 257
12. M. Tiglio, L. Lehner, D. Neilsen: 3D simulations of Einstein's equations: symmetric hyperbolicity, live gauges and dynamic control of the constraints. `gr-qc/0312001` 257, 259
13. H. Friedrich, G. Nagy: Commun. Math. Phys. **201**, 619 (1999) 270
14. H.-O. Kreiss, O.E. Ortiz: Lect. Notes Phys. **604**, 359 (2002) 270
15. B. Szilágyi, B. Schmidt, J. Winicour: *Phys. Rev.* D **65**, 064015 (2002) 270
16. C. Bona, C. Palenzuela: Lect.Notes Phys. **617**, 130 (2003) 259

Index

Lecture Notes in Physics

For information about earlier volumes
please contact your bookseller or Springer
LNP Online archive: springerlink.com

Vol.643: F. Strocchi, Symmetry Breaking

Vol.644: B. Grammaticos, Y. Kosmann-Schwarzbach, T. Tamizhmani (Eds.) Discrete Integrable Systems

Vol.645: U. Schollwöck, J. Richter, D. J. J. Farnell, R. F. Bishop (Eds.), Quantum Magnetism

Vol.646: N. Bretón, J. L. Cervantes-Cota, M. Salgado (Eds.), The Early Universe and Observational Cosmology

Vol.647: D. Blaschke, M. A. Ivanov, T. Mannel (Eds.), Heavy Quark Physics

Vol.648: S. G. Karshenboim, E. Peik (Eds.), Astrophysics, Clocks and Fundamental Constants

Vol.649: M. Paris, J. Rehacek (Eds.), Quantum State Estimation

Vol.650: E. Ben-Naim, H. Frauenfelder, Z. Toroczkai (Eds.), Complex Networks

Vol.651: J. S. Al-Khalili, E. Roeckl (Eds.), The Euroschool Lectures of Physics with Exotic Beams, Vol.I

Vol.652: J. Arias, M. Lozano (Eds.), Exotic Nuclear Physics

Vol.653: E. Papantonoupoulos (Ed.), The Physics of the Early Universe

Vol.654: G. Cassinelli, A. Levrero, E. de Vito, P. J. Lahti (Eds.), Theory and Appplication to the Galileo Group

Vol.655: M. Shillor, M. Sofonea, J. J. Telega, Models and Analysis of Quasistatic Contact

Vol.656: K. Scherer, H. Fichtner, B. Heber, U. Mall (Eds.), Space Weather

Vol.657: J. Gemmer, M. Michel, G. Mahler (Eds.), Quantum Thermodynamics

Vol.658: K. Busch, A. Powell, C. Röthig, G. Schön, J. Weissmüller (Eds.), Functional Nanostructures

Vol.659: E. Bick, F. D. Steffen (Eds.), Topology and Geometry in Physics

Vol.660: A. N. Gorban, I. V. Karlin, Invariant Manifolds for Physical and Chemical Kinetics

Vol.661: N. Akhmediev, A. Ankiewicz (Eds.) Dissipative Solitons

Vol.662: U. Carow-Watamura, Y. Maeda, S. Watamura (Eds.), Quantum Field Theory and Noncommutative Geometry

Vol.663: A. Kalloniatis, D. Leinweber, A. Williams (Eds.), Lattice Hadron Physics

Vol.664: R. Wielebinski, R. Beck (Eds.), Cosmic Magnetic Fields

Vol.665: V. Martinez (Ed.), Data Analysis in Cosmology

Vol.666: D. Britz, Digital Simulation in Electrochemistry

Vol.667: W. D. Heiss (Ed.), Quantum Dots: a Doorway to Nanoscale Physics

Vol.668: H. Ocampo, S. Paycha, A. Vargas (Eds.), Geometric and Topological Methods for Quantum Field Theory

Vol.669: G. Amelino-Camelia, J. Kowalski-Glikman (Eds.), Planck Scale Effects in Astrophysics and Cosmology

Vol.670: A. Dinklage, G. Marx, T. Klinger, L. Schweikhard (Eds.), Plasma Physics

Vol.671: J.-R. Chazottes, B. Fernandez (Eds.), Dynamics of Coupled Map Lattices and of Related Spatially Extended Systems

Vol.672: R. Kh. Zeytounian, Topics in Hyposonic Flow Theory

Vol.673: C. Bona, C. Palenzula-Luque, Elements of Numerical Relativity

Vol.674: A. G. Hunt, Percolation Theory for Flow in Porous Media

Vol.675: M. Kröger, Models for Polymeric and Anisotropic Liquids

Vol.676: I. Galanakis, P. H. Dederichs (Eds.), Half-metallic Alloys

Vol.678: M. Donath, W. Nolting (Eds.), Local-Moment Ferromagnets

Vol.679: A. Das, B. K. Chakrabarti (Eds.), Quantum Annealing and Related Optimization Methods

Vol.680: G. Cuniberti, G. Fagas, K. Richter (Eds.), Introducing Molecular Electronics

Vol.681: A. Llor, Statistical Hydrodynamic Models for Developed Mixing Instability Flows

Vol.682: J. Souchay (Ed.), Dynamics of Extended Celestial Bodies and Rings

Vol.683: R. Dvorak, F. Freistetter, J. Kurths (Eds.), Chaos and Stability in Planetary Systems

Vol.685: C. Klein, O. Richter, Ernst Equation and Riemann Surfaces

Vol.686: A. D. Yaghjian, Relativistic Dynamics of a Charged Sphere

Vol.687: J. W. LaBelle, R. A. Treumann (Eds.), Geospace Electromagnetic Waves and Radiation

Vol.688: M. Rubi, M. C. Miguel (Eds.), Jamming, Yielding, and Irreversible Deformation in Condensed Matter

Vol.689: W. Pötz, J. Fabian, U. Hohenester (Eds.), Quantum Coherence

Vol.691: S.S. Abdullaev, Construction of Mappings for Hamiltonian Systems and Their Applications

Vol.692: J. Frauendiener, D.J.W. Giulini, V. Perlick (Eds.), Analytical and Numerical Approaches to Mathematical Relativity