FOUNDATIONS OF GENERAL RELATIVITY

KLAAS LANDSMAN

FROM EINSTEIN TO BLACK HOLES

RADBOUD
UNIVERSITY
PRESS

FOUNDATIONS OF GENERAL RELATIVITY

FROM EINSTEIN TO BLACK HOLES

KLAAS LANDSMAN

Foundations of General Relativity: From Einstein to Black Holes

Published by RADBOUD UNIVERSITY PRESS
Postbus 9102, 6500 HC Nijmegen, the Netherlands
www.radbouduniversitypress.nl | www.ru.nl/radbouduniversitypress
radbouduniversitypress@ru.nl

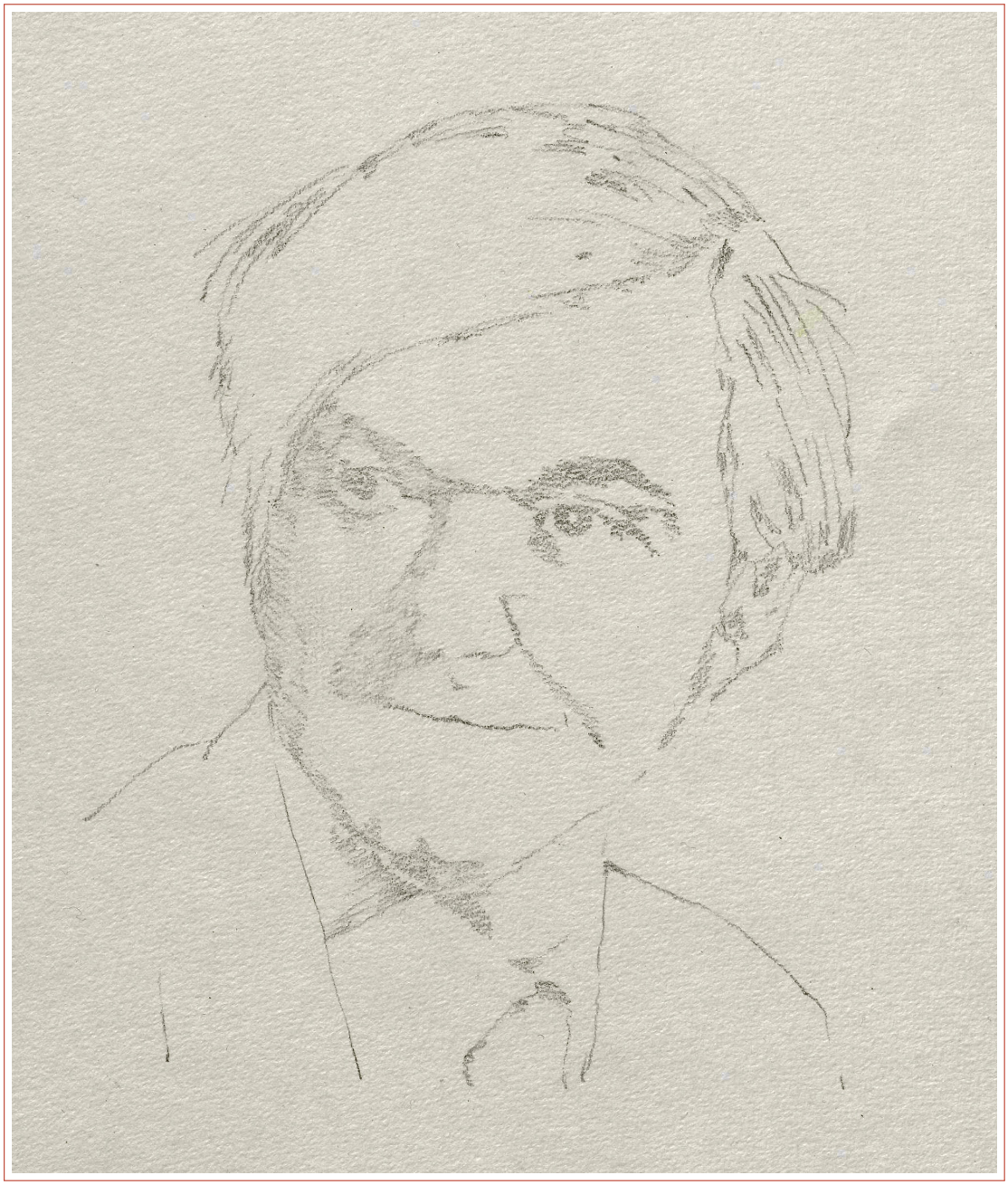
Cover image: © Edith de Jong
Cover design: Pumbo.nl
Print and distribution: Pumbo.nl
Version: 2021-09

ISBN: 978 90 831 789 29
DOI: http://doi.org/10.54195/EFVF4478
Free download at: www.radbouduniversitypress.nl

Institute for Mathematics, Astrophysics, and Particle Physics
Radboud University, Nijmegen, The Netherlands

RADBOUD
UNIVERSITY
PRESS

Dedicated to Roger Penrose



Preface

This book grew out of lecture notes for my master's courses on general relativity (GR) and on singularities and black holes taught at Radboud University (Nijmegen). These notes were originally intended for our students with a double bachelor degree in mathematics and physics, but in its final form the book is intended for *all* students of GR of *any* age and orientation who have a background including at least first courses in special and general relativity, differential geometry, and topology.[1] The recent textbook *Elements of General Relativity* by Chruściel (2019) would make students singularly well prepared for this one, but almost any introduction to GR, combined with the typical mathematical background in manifolds etc. that is usually included in such introductions, will do. This book, then, is a *second, mathematically oriented course in general relativity*, with extensive references and occasional excursions in the history and philosophy of gravity, including a relatively lengthy historical introduction. As such, it omits standard physics material like the classical tests etc. Furthermore, the material is developed in such a way that through the last two chapters the reader may acquire a taste of the modern mathematical study of black holes initiated by Penrose, Hawking, and others, so that successful readers might be able to begin reading research papers in this direction, especially in mathematical physics and in the philosophy of physics. This focus comes with an introduction to what is called *causal theory*, but alas, it also implies that in order to keep the book medium-sized I had to omit applications like cosmology and gravitational waves. In any case I hope the book appeals to mathematicians, physicists, and philosophers–perhaps even historians–of physics alike.

My own experience is that a really deep field such as GR (or quantum theory) can only be learned by *reading* a large number of books saying the right things in different ways, as well as by *talking* to good people working in the field. As a reader, my first encounter with GR was Einstein's own exposition *Relativity: The Special and General Theory* (Einstein, 1921), which is still in print. In the summer of 1981, having just graduated from highschool, this was followed by two books that were a little more difficult, namely *Space - Time - Matter* by Weyl (1922) and *The Mathematical Theory of Relativity* by Eddington (1923), both of which are not only highly mathematical but also profoundly philosophical in spirit. Weyl makes this point himself:

> At the same time it was my wish to present this great subject as an illustration of the intermingling of philosophical, mathematical, and physical thought, a study which is dear to my heart. This could only be done by building up the theory systematically from the foundations and by restricting attention throughout to the principles. But I have not been able to satisfy these self-imposed requirements: the mathematician predominates at the expense of the philosopher.[2] (Weyl, 1918, Preface)

Indeed, Weyl, Eddington and Einstein were *natural philosophers* in the spirit of the scientific revolution, whose mix of physics, mathematics, and philosophy was the key to its success. Hence it seems hardly a coincidence that Einstein was Newton's successor, for if *any* scientific theory has ever represented the *Philosophiae Naturalis Principia Mathematica*, it must be GR.

[1] Logically speaking, the GR material is even developed from scratch, and indeed the first course in this direction that I taught was optimistically offered also to mathematics students without any physics background. But experience shows that the material makes little sense without some prior exposure to both special and general relativity.

[2] 'Zugleich wollte ich an diesem Großem Thema ein Beispiel geben für die gegenseitige Durchdringung philosophischen, mathematischen und physikalischen Denkens, die mir sehr am Herzen liegt; dies konnte nur durch einen völlig in sich geschlossenen Aufbau von Grund auf gelingen, der sich durchaus auf das Prinzipielle beschränkt. Aber ich habe meinen eigenen Forderungen in dieser Hinsicht nicht voll Genüge tun können: der Mathematiker behielt auf Kosten des Philosophen das Übergewicht.' Translation: Henry L. Brose (Weyl, 1922).

So even though at the time I understood almost nothing of the technical content of their books, Einstein, Weyl, and Eddington left an indelible mark in the way they approached natural science through mathematics and philosophy. Still during that same long summer vacation between highschool and university in 1981, which I regard as one of the high points of my life, I also bought *Gravitation* by Misner, Thorne & Wheeler (1973). For a while I considered this the greatest book written on any topic whatsoever,[3] and when Misner, well in his eighties at the time, not only came to a talk I gave in one of Bub's *New Directions in the Foundations of Physics* conferences in Washington DC but even asked a question, after he had answered positively to my counter-question if he was Charles Misner I was petrified and unable to say anything.[4]

My next book was *The Large Scale Structure of Space-Time* by Hawking & Ellis (1973), and so on, until *General Relativity and the Einstein Equations* by Choquet-Bruhat (2009) and most recently *The Geometry of Black Holes* by Chruściel (2020). These are all masterpieces written by founders of the field; like most students and authors in mathematical GR I am also indebted to Penrose (1972), O'Neill (1983) and Wald (1984). Furthermore, Earman (1995) set the stage in the philosophy of physics. Other influences on this text include Weinberg (1972), Kriele (1999), Poisson (2004), Schoen (2009), Gourgoulhon (2012), Malament (2012), and Minguzzi (2019).

This brings me to the question why an author who so far wrote little on GR is entitled to write a book about the subject–even if it has been an almost lifelong passion. In the first of the *Jeeves and Wooster* episodes (about an indolent English aristocrat and his butler), Wooster's aunt asks:

> Do you work, Mr Wooster?

upon which Wooster (i.e. the aristocrat), taken aback by her question, mumbles:

> Well, I've known a few people who work.

I've known a few people who work, too (in GR, that is). The greatest of these, in my view, is Roger Penrose, to whom this book is dedicated in honour of his pivotal role in the creation of mathematical relativity and the modern theory of singularities and black holes,[5] combined with a singular lack of pomp and circumstance, for a scientist of his calibre. In her recent autobiography, Yvonne Choquet-Bruhat, who has known Penrose for over 50 years, puts it well:

> In spite of his successes, he remains a man without pretension, open and friendly. He came to listen, a few years ago, to a talk I gave at a seminar in Oxford. Afterwards we had lunch with a few colleagues and the conversation turned to the publication of his complete works. Penrose said: 'My problem is to know if I must correct my mistakes before publication.' It is a great quality to recognize a mistake, even small. Few human beings, scientists or not, are ready to do it. (Choquet-Bruhat, 2018, chapter 10)

Perhaps the key to his success, which on the one hand seems typical for most great scientists and artists but on the other hand seems paradoxical as a path to influence and eminence, is this:

[3] Kaiser (2012) gives an interesting perspective on *Gravitation* and its history, which confirms its uniqueness.

[4] Nonetheless, I now see a basic drawback of *Gravitation*: with its xxvi + 1279 pages, *it leaves no room for the reader* (except in doing the exercices, which I all duly did in the next few years), who is overwhelmed and cornered.

[5] A scientific biography of Penrose remains to be written (in 2019 Dennis Lemkuhl conducted a series of interviews with Penrose). For now, see e.g. Thorne (1994), Frauendiener (2000), Friedrich (2011), and Ellis (2014). Both the written AIP interview by Lightman (1989) and the videotaped interview by Turing's biographer and Penrose's former student Hodges (2014) are great and intimate portraits of Penrose.

> It was important for me always, if I wanted to work on a problem, to think I had a different angle on it from other people. Because I wasn't good at following where everybody else went. I wasn't the kind of person who could pick up the prevalent arguments and knowledge of the time. Other people were good at that. They could suck it all out and put it together and make advances. I was the kind of person who'd have some kind of quirky way of looking at something on my own, which I would hide away and work at. So it meant that I had to have some way of looking at a problem that was my own.[6]

Here one would like to emphasize that the word 'looking', used twice, should be taken quite literally: as he also emphasized elsewhere, Penrose is primarily a *visual thinker*. This is exemplified most famously by his invention of the diagrams named after him, but it goes back a long way, including for example the "impossible figures" he created with his father, and his interaction with the Dutch artist Maurits Cornelis Escher (1898–1972).[7] Penrose usually drew his own figures in a professional, yet playful and characteristic way, and each of them not only makes some scientific point but is also a pleasure to look at. A few are reproduced in this book.

I first heard Penrose speak in Cambridge in 1989 about his recent book *The Emperor's New Mind: Concerning Computers, Minds and The Laws of Physics*, later superseded by the equally controversial *Shadows of the Mind: A Search for the Missing Science of Consciousness* (1994). I got to know him personally during a *Seven Pines Symposium* in Minnesota in 2005, where the organizers had the luminous idea that famous and ordinary participants share an apartment. I am not sure to which one of us this arrangement was initially more shocking, but we got along well, and he very kindly came to the opening conference of our institute IMAPP at Nijmegen (2005) as a speaker (forming part of a stellar line-up including for example physicist Gerard 't Hooft, mathematician Don Zagier, and theologian Hans Küng),[8] where he explained the key ideas of his later book *Cycles of Time* (2010). Having him all for myself for 1.5 hours, I then drove him to the famous *Amstel Hotel* in Amsterdam, the most expensive hotel in the country, since I felt that if that is the place where Bob Dylan and the like stay, certainly also Roger belonged there.[9] He later returned to the Netherlands for a mathematical physics conference and usually came to my talks when I was visiting Oxford and joined for lunch or dinner whenever possible.

Dominating the public image, Stephen Hawking was unquestionably another key figure in mathematical relativity.[10] I observed Hawking on an almost daily basis between 1989–1997, when I was a postdoc at DAMTP in Cambridge, but I wasn't in his group and never talked to him directly. I did mingle with his circle though, and inhaled a certain culture from this. Although in the wake of his *Brief History of Time* (1988) Stephen had by then become a scientific superstar, it is only after his death in 2018 that I really came to appreciate his genius and his life.[11]

Hence this book has been heavily influenced by Hawking and Penrose, and of course includes their singularity (i.e. incompleteness) theorems, but without being blind to other developments, notably the initial-value or PDE approach to GR, which, as will be explained in detail especially in connection with cosmic censorship, sometimes leads to a different perspective on space-time.

[6] Quoted from the AIP interview by Lightman (1989).

[7] See Wright (2014) for the history of Penrose diagrams; Wright (2013) explains the link with Escher. Penrose admired Escher at least since 1954, when, as a student participant to the *International Conference of Mathematicians* in Amsterdam, he saw an Escher exhibition. In 1962 Penrose visited Escher at his home in Baarn. See Penrose (2005), chapter 2, and the TV documentary Penrose (2015). See also §1.9 and footnote 486.

[8] At the conference dinner we gave each speaker an expensive Japanese wooden puzzle, which we asked them to solve as quickly as possible. Penrose won easily (which, in good spirits, greatly annoyed 't Hooft and Zagier).

[9] Our financial staff did not appreciate this and I paid for his room, with a river view doubling the prize, myself.

[10] See §1.9 for some brief historical comments on the development of mathematical GR.

[11] See Hawking (1999) for an unusually honest account of this life; the second (2007) edition is milder.

A complete coverage of the causal theory is both impossible in a work of this size and undesirable for students looking for a first encounter, but fortunately it is also unnecessary in view of the recent encyclopedic (Open Access) treatment by Minguzzi (2019), which always lies open on my desk. Similarly, a complete description of the PDE approach to GR would require not only a very different author (or rather a team of authors), but also much more space including preliminary material. So we are fortunate to have Ringström (2009) for those who want more than the *very* first introduction given here, as well as Klainerman & Nicolò (2003) and Christodoulou (2008). I have tried to do justice to the modern spirit of mathematical relativity, which is characterized by a mix of the causal and the PDE theories and culminates in the cosmic censorship and final state conjectures. The aim of this book is not at all to describe the latest news about such matters, but merely to explain what the discussions are about, and give students and more senior readers not specializing in this area and entry point to the research literature. Likewise for the no-hair or uniqueness theorems for black holes and black hole thermodynamics, with which the book ends. Thus the book stops not only where (mathematical or philosophical) research papers on *classical* GR begin, but also where *quantum* aspects of gravity begin.

Finally, as may be expected more from a work in the humanities than in mathematical physics (between which the history and philosophy of physics resides), there are almost 700 footnotes, placed where the name "footnote" suggests they belong. They contain credits (e.g. for some of the arguments and derivations I give) and other pointers to the literature, as well as additional information that refines or qualifies the mathematics just discussed, and/or adds conceptual or historical information I found interesting. They may be skipped in principle by those who just want to hear the melody, but they seem to me to be essential for enjoying the full sound.

For a more detailed summary of this book the prospective reader is encouraged to take a look at the synopsis and the table of contents, which in this order immediately follow this preface.

I received very kind help and feedback from a number of students and colleagues, of whom I would like to mention Ibai Asensio Pol, Jeremy Butterfield, Erik Curiel, Jeroen van Dongen, Juliusz Doboszewski, John Earman, Jan Głowacki, Evert-Jan Hekkelman, Leo Garcia Heveling, Michel Janssen, Dennis Lemkuhl, Martin Lesourd, Sera Markoff, Ettore Minguzzi, John Norton, Bryan Roberts, Quinten Rutgers, and Jan Sbierski. Most chapters were also reviewed during the 2020–2021 Cambridge–LSE *Philosophy of Physics Bootcamp*, which was of great help.

The final edit of this book was done during July 2021 at a lovely cottage by the river IJssel, which we could use thanks to the generous hospitality of our friends Arend and Esther van der Sluis. This last round also benefited from the online conference *Singularity theorems, causality, and all that: A tribute to Roger Penrose*, held in June 2021 (organized by Piotr Chruściel, Greg Galloway, Michael Kunzinger, Ettore Minguzzi, and Roland Steinbauer) where I could pick up the latest news and was also given the unexpected honour to speak. My greatest debt, however, is to Edith de Jong, who contributed so much more than the beautiful cover art and various drawings to this book, including the one of Penrose on the dedication page of this book.[12]

[12]This drawing is based on a photograph of Penrose in *Gravitation* (Misner, Thorne, and Wheeler, 1973, p. 936).

Synopsis

Here is a brief summary of the chapters, which may also help potential readers as well as instructors. My own experience is that chapters 2, 3, 4, 5, and 7 may form the basis of a one-semester (master's) course entitled *Mathematical structure of general relativity*.[13] This may be followed by another one-semester course called *Singularities and black holes*,[14] based on chapters 6, §§8.1–8.4, 9, and a selection of topics from chapter 10. For advanced students with sufficient background in both GR and mathematics, the latter could also stand alone.[15]

Since this book is also, perhaps even largely, intended for self-study and pleasure, it contains no exercises. However, instructors (and even students with enough self-discipline) can easily assign almost any derivation as an exercise (for themselves). Many difficult results are just mentioned without proof (always with a reference), and these could serve as advanced problems.

1. *Historical introduction.* Based on recently completed research by historians of science, the reader is introduced to Einstein's "bumpy road" to his theory of general relativity. Although GR may well be the most sublime of all scientific theories, created by a man who is widely–perhaps exaggeratedly–seen as one of the supreme geniuses humanity has brought about, at least the story of its discovery is "human, all too human". I also include a little mathematical history, involving Riemann and others, as well as a brief picture of mathematical GR until about 1970. I close with some musings on general covariance.

2. *General differential geometry.* This is a turbo introduction to manifolds and tensors, intended for readers who have already seen some basic treatment of this material. Even within some modern, coordinate-free approach to differential geometry, both abstract and computational aspects of GR also require the use of old-fashioned coordinates and indices.

3. *Metric differential geometry.* Here the pace slows down. In this brief chapter, which is mainly a warm-up for the next two chapters, metrics, geodesics, connections, and the Levi-Civita (i.e. metric) connection are introduced. This material is totally standard, but I have done my best to give some perspective on geodesics in Lorentzian manifolds.

4. *Curvature.* This chapter may have an unusual emphasis on sectional curvature, constant curvature, and the nineteenth century origins of the abstract modern theory in submanifolds of Euclidean space. In my experience, this background is especially helpful in understanding the Gauss–Weingarten and Gauss-Codazzi equations, which in turn are essential for the derivation of the constraint equations of GR. In the same spirit, the last section discusses the classical "fundamental theorem for hypersurfaces", which gives necessary and sufficient conditions for the existence and (geometric) uniqueness of embeddings of curved surfaces in flat space. Though much simpler, this theorem resembles the corresponding result for the Einstein equations in §7.6, notably regarding the role of constraints.

5. *Geodesics and causal structure.* This chapter introduces the topological and geometric techniques, largely developed by Penrose and others in the 1960s, that demarcate *mathematical* GR from a *theoretical physics* treatment. As already mentioned, our discussion is far from complete, but it is hopefully enough to advocate a specific *causal* way of thinking.

[13] Chapter 2 should perhaps not be discussed in detail (which might repel students); it alone has a summary (§2.7).

[14] Nonetheless, putting chapter 6 before chapter 7 in this book is a logical choice, since the singularity theorems do not rely on the Einstein equations. The second half of chapter 8 contains advanced and partly speculative "retro" material that I simply find interesting–especially the unresolved problem of time–and could not resist including.

[15] Natário (2021), which I saw much too late to use it, provides a one-semester course in all of mathematical GR.

The ensuing causal theory has turned out to be very fruitful for the modern study of both black holes and PDE aspects of GR. In both respects, one of the central ideas in all of mathematical GR is *global hyperbolicity*. This concept is studied from several equivalent definitions, which have been brought to completion only in the twenty-first century.

6. *The singularity theorems of Hawking and Penrose.* Penrose's singularity theorem from 1965–which should more aptly be called an *incompleteness* theorem–remains the most powerful illustration of the techniques of the previous chapter. But for pedagogical reasons I start with Hawking's singularity theorem (*idem dito*), which postdated Penrose's but is easier since it does not involve the often counterintuitive "lightlike" (or "null") reasoning that is typical of Penrose's theorem (and indeed of almost all of his work in GR). The opening pages of the chapter also provide some insight into the struggle of finding an appropriate definition of space-time singularities, from Einstein to Penrose and Hawking.

7. *The Einstein equations.* In standard fashion, the Einstein (field) equations are derived from an action principle, with extra attention however for boundary terms. This also involves a brief treatment of matter sources (i.e. the energy-momentum tensor). The main goal of the chapter is to introduce the specific PDE analysis of the Einstein equations introduced by Choquet-Bruhat in the 1950s, which she completed in a joint paper with Geroch from 1969. This analysis provides and solves a geometric initial-value formulation for the Einstein equations, which is far from obvious and circumvents all kinds of conceptual and technical questions involving equations posed on a space-time that does not (yet) exist.

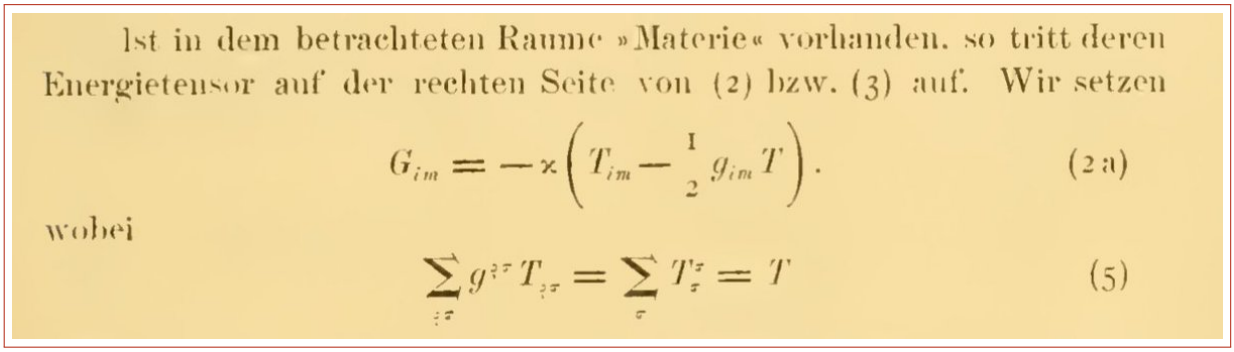
8. *The 3+1 split of space-time.* For both technical and conceptual reasons–we do not experience space-time but space and time–it is often helpful to take a "non-relativistic" view on the Einstein equations. This involves an arbitrary foliation of space-time into spacelike hypersurfaces, controlled by the lapse and shift functions of the physics literature (i.e. the ADM formalism). Thus the Einstein equations are cleanly split into propagation equations and constraint equations, and one has an easy transition to a Hamiltonian formalism. The last section concerns the deceptive "problem of time", which is more or less debunked.

9. *Black holes I: Exact solutions.* The theory of black holes is an interplay between abstract arguments, like Penrose's singularity theorem and associated techniques, and concrete examples. This chapter is devoted to the latter. After a warm-up on de Sitter space (which is not singular but has some kind of horizon), which may be skipped, we study the three main cases of interest: Schwarzschild (including the Kruskal extension), Reissner–Nordström, and Kerr. Especially the latter is a source of endless fascination, which can only be sparked.

10. *Black holes II: General theory.* Penrose remains a central figure in the model-independent study of black holes, e.g. through his four closely related concepts of conformal completion, null infinity, (absolute) event horizon (which leads to a mathematical definition of a black hole), and the diagram named after him; by means of cosmic censorship, and via the Penrose inequality. Furthermore, he unearthed the structure of various black hole horizons (namely event horizons, Cauchy horizons, and Killing horizons, in introducing all of which also Hawking played a major part) as null hypersurfaces ruled by lightlike geodesics. The last two sections on uniqueness or "no hair" theorems and on thermodynamics of black holes are introductory; alas, they merely scratch the surface of these miraculous topics.

Finally, Appendix A on *Lie groups, Lie algebras, and constant curvature* mainly supports §4.4, whereas Appendix B on *Formal PDE theory* gives some background for especially §7.6.

Contents



Ist in dem betrachteten Raume »Materie« vorhanden, so tritt deren Energietensor auf der rechten Seite von (2) bzw. (3) auf. Wir setzen

$$G_{im} = -\varkappa\left(T_{im} - \frac{1}{2} g_{im} T\right). \qquad (2\,a)$$

wobei

$$\sum_{i\sigma} g^{i\sigma} T_{i\sigma} = \sum_{\sigma} T^{\sigma}_{\sigma} = T \qquad (5)$$

1 Historical introduction

On 25 November 1915, Einstein submitted a paper containing the above equations, which (in an appropriate mathematical context) state his general theory of relativity (GR). Einstein thereby replaced Newton's theory of universal gravity from 1687, and in 1919 he became famous overnight when the historical expedition led by Eddington confirmed Einstein's prediction–against Newton–of the gravitational deflection of starlight passing near the Sun. Einstein also computed the correct perihelion shift of Mercury, which had been known since 1859 as an anomaly in Newtonian gravity. Still within his conceptual reach were, for example, the properties of the binary pulsar PSR B1913+16, discovered in 1974 by Hulse and Taylor, as well as the gravitational waves detected by the LIGO experiment in 2015, almost a century after Einstein had predicted their existence. Beyond what Einstein himself foresaw or could bear, GR also turned out to describe the expansion of the cosmos and hence–unless quantum theory intervenes–its origin in a big bang (Einstein initially denied the former and never accepted the latter implication).[16]

Last but not least, GR suggests the possibility of black holes (another "singular" phenomenon allowed by his theory that Einstein stubbornly kept disavowing), and, in the hands of Penrose, gives compelling conditions for their existence. The fact that these conditions are met in the universe is now beyond any (astrophysical) doubt, as reconfirmed by the spectacular image of the supermassive black hole M87* revealed in 2019 by the Event Horizon Telescope (EHT).

Hilbert, the greatest mathematician of his time, expressed his admiration for GR as follows:

> Constructing the theory of general relativity is, in my opinion, one of the greatest achievements in scientific history. The edifice that Pythagoras started and Newton continued has been completed by Einstein.[17] (Hilbert, 1920)

[16] With hindsight the occurrence of a big bang is implicit in the specific solution to Einstein's equations that describes an expanding homogeneous and isotropic universe, found independently by Friedman in 1922 and Lemaître in 1927. The latter also matched this with contemporary observations of redshifts of galaxies (often but not quite rightly attributed to Hubble), and is the originator of the physical idea of a hot early state of the universe, which he proposed in the early 1930s. See e.g. Kragh (2007) and Nussbaumer & Bieri (2009). In 1965 the 2.7K Cosmic Microwave Background was discovered (by coincidence) by Penzias and Wilson. The CMB was interpreted almost at once as a relic of the big bang by Dicke and Peebles and others, which matched and revived earlier calculations by Gamow and others of the abundances of hydrogen and helium in stars. Within this *Zeitgeist* Hawking's singularity theorem from 1966, which we will discuss in detail, gives the final *mathematical* underpinning of the big bang.

[17] 'Die Aufstellung der allgemeinen Relativitätstheorie ist m.E. eine der größten Leistungen in der Geschichte der Wissenschaften. Den von Pythagoras begonnenen, von Newton ausgestalteten, Bau hat Einstein zum Abschluß gebracht.' Quoted in Corry (1999, p. 522) from unpublished lecture notes from Hilbert's 1920 course *Mechanik und neue Gravitationstheorie*. As we shall see in §1.7, the relationship between Einstein and Hilbert had ups and downs.

Hilbert knew what he was talking about; we will return to his role in the history of GR (see §1.7), and more generally to the profound interaction between physics and mathematics in this theory. In 1918 Hilbert's pupil Weyl, a contemporary of Einstein's and later colleague of his at the Institute for Advanced Study in Princeton, wrote the first textbook about GR, called *Raum - Zeit - Materie (Space - Time - Matter)*. Its preface starts in the following, even more lyrical way:

> Einstein's Theory of Relativity has advanced our ideas of the structure of the cosmos a step further. It is as if a wall which separated us from Truth has collapsed. Wider expanses and greater depths are now exposed to the searching eye of knowledge, regions of which we had not even a presentiment. It has brought us much nearer to grasping the plan that underlies all physical happening.[18] (Weyl, 1918, Vorwort)

Perhaps Einstein himself was not entirely neutral about his work, but he was as lyrical:

> The theory has unsurpassed beauty (...) My boldest dreams have been fulfilled (...) That I was given to experience this (...) The highest satisfaction of my life.[19]

It is all the more remarkable that Einstein found his theory under pretty miserable circumstances. First, Germany was in a war which he was one of the very few people (on either side) to oppose. This isolated him even among his colleagues, who in any case hardly understood his scientific quest. Second, he was separated from first wife (Mileva) and their two sons (Hans Albert and Eduard).[20] Einstein's ability to not only continue his work under such conditions but even produce one of the greatest scientific theories of all times was later explained as follows:

> His true passion lay in the understanding of the riddle of the immeasurable world, which stood outside and above the bickering and wriggling of personal interests, feelings, and urges of people. Seeking such understanding comforted him from the moment he had seen through the hypocrisy of the common ideals of decency. The contemplation of this external reality lured him like a liberation from an earthly prison.[21] (Fokker, 1955)

Einstein's construction of GR was the culmination of an epic quest for the structure of space and time, which had started with his special theory of relativity from 1905. Unlike his earlier relativity theory, Einstein's road to GR is extremely well documented.[22] The summary that now follows suggests that the key to Einstein' success was not some superhuman genius à la Newton but his ability to recognize inconsistencies (including his own mistakes) and take it from there.

---

[18] 'Mit der Einsteinschen Relativitätstheorie hat das menschliche Denken über den Kosmos eine neue Stufe erklommen. Es ist, als wäre plötzlich eine Wand zusammengebrochen, die uns von der Wahrheit trennte: nun liegen Weiten und Tiefen vor unserm Erkenntnisblick entriegelt da, deren Möglichkeit wir vorher nicht einmal ahnten. Der Erfassung der Vernunft, welche dem physischen Weltgeschehen innewohnt, sind wir einen gewaltigen Schritt näher gekommen. ' Translation: Henry L. Brose, from the fourth edition from 1922 (see also §1.8).

[19] 'Die Theorie ist von unvergleichbarer Schönheit (...) Ich war einige Tagen fassungslos von Erregung (...) Die kühnsten Träume sind nun in Erfüllung gegangen (...) Dass ich das habe erleben dürfen (...) Die höchste Befriedigung meines Lebens'. These quotations can be found in the original German in Fölsing (1993), chapter 4.

[20] On the other hand, whilst officially still married to Mileva he had started a relationship with Elsa Einstein, who also lived in Berlin (Einstein was in fact her maiden name; she was both a first and second cousin to Albert Einstein). They got married in 1919 after Einstein's divorce from Mileva and stayed together until Elsa's death in 1936.

[21] Dutch original: 'Zijn ware hartstocht lag in het doorgronden van het raadsel der onmetelijke wereld, die buiten en boven het geharrewar en het gewriemel van persoonlijke belangen, gevoelens, en driften der mensen stond. Dat nadenken troostte hem toen hij de schijnheiligheid van de gangbare fatsoenlijke idealen had doorzien. Als een bevrijding uit een aardse gevangenis lokte hem de beschouwing van die buitenpersoonlijke werkelijkheid.'

[22] The main sources for the history of GR are Einstein (1996ab) and Renn (2007), the massive scholarship in which is based largely on the work of Michel Janssen, John Norton, Jürgen Renn, Tilman Sauer, and John Stachel. See also van Dongen (2010, 2017), Janssen (2014), and Janssen & Renn (2020). Standard biographies are Fölsing (1993) and Isaacson (2017). The only scientific biography of Einstein (Pais, 1982) is now largely outdated.

1.1 From physical principles to a mathematical framework

In an illuminating informal lecture he gave on 14 December 1922 to students at Kyoto University, Einstein recalled his first steps towards what ultimately became general relativity:

> The first thought leading to the general theory of relativity occurred to me two years later, in 1907, and it did in a memorable setting. I was already dissatisfied with the fact that the relativity of motion is restricted to motion with constant relative velocity and does not apply to arbitrary motion. I had always wondered privately whether this restriction could somehow be removed. In 1907, while trying, at the request of Mr. Stark, to summarize the results of the special theory of relativity for the *Jahrbuch der Radioaktivität und Elektronik* of which he was the editor, I realized that, while all other laws of nature could be discussed in terms of the special theory of relativity, the theory could not be applied to the law of universal gravitation. I felt a strong desire to somehow find out the reason behind this. But this goal was not easy to reach. What seemed to me most unsatisfactory about the special theory of relativity was that, although the theory beautifully gave the relationship between inertia and energy, the relationship between inertia and weight, i.e., the energy of the gravitational field, was left completely unclear. I felt that the explanation could probably not be found at all in the special theory of relativity. I was sitting in a chair in the Patent Office in Bern when all of a sudden I was struck by a thought: "If a person falls freely, he will certainly not feel his own weight." I was startled. This simple thought made a really deep impression on me. My excitement motivated me to develop a new theory of gravitation. My next thought was: "When a person falls, he is accelerating. His observations are nothing but observations in an accelerated system." Thus, I decided to generalize the theory of relativity from systems moving with constant velocity to accelerated systems. I expected that this generalization would also allow me to solve the problem of gravitation. This is because the fact that a falling person does not feel his own weight can be interpreted as due to a new additional gravitational field compensating the gravitational field of the Earth, in other words, because an accelerated system gives a new gravitational field.[23]

This recollection makes, and somewhat conflates or refers to, at least three different points:

1. Einstein was haunted by the idea that the "principle of relativity"–which in his special theory (as well as in Newtonian mechanics) only applies to motion with constant velocity–should be extended to arbitrary motion, in that the laws of physics should be the same in *any* frame of reference.[24] Since the special principle of relativity makes uniform motion relative, Einstein called his extended version the "general principle of relativity", which he considered so important that he would later even name his entire theory after it.

[23] The English translation of the original notes in Japanese taken by Einstein's tour guide Yun Ishiwara comes from Einstein (2013), pp. 637–638. Another translation may be found in *Physics Today*, August 1932, p. 45.

[24] If true, this would make all kinds of motion in (otherwise) empty space indistinguishable, and hence both inertial and accelerated motion would effectively be undefined (as opposed to the situation in both Newtonian mechanics and, indeed, GR). As a way out, Mach proposed that motion is exclusively defined with respect to all (other) matter in the universe, even if this only consists of distant stars. Following Einstein himself, this is often called ***Mach's principle***, although Barbour (1989, *Introduction*), claims that Einstein misunderstood Mach's idea by conflating its application to Newton's first law (for which it was apparently intended) and his second (for which it was not). In any case, Einstein was initially guided by this principle, which until 1918 he did not clearly distinguish from both the equivalence principle and the principle of general covariance. Although Mach's principle fails in GR, as Einstein gradually came to recognize after 1920 (for example, non-flat vacuum solutions to the Einstein equations violate it), it nonetheless helped Einstein significantly in his search for the field equations (Janssen, 2014).

2. Newton's theory of gravity, based on action at a distance, was not only physically absurd (as Newton had already noted himself), but also stood in an uneasy relationship (to say the least) with special relativity, which postulates the velocity of light as the largest signal velocity ('the theory could not be applied to the law of universal gravitation').

3. Finally, the mysterious equality of gravitational and inertial mass (which in Newtonian physics is a curious coincidence) led Einstein to introduce the following two-sided coin:

   (a) Freely falling observers, who according to someone at rest are accelerating, feel no gravity: they may even consider themselves at rest as if there were no gravity.

   (b) Accelerating observers in a situation without gravity may equally well consider themsleves at rest in a specific gravitational field (pulling in the opposite direction).

Point 1 is controversial and warrants further discussion; see §1.10. Point 2 is resolved by GR; for example, test particles respond to the *local* structure of space-time by moving on geodesics.[25] Part 3(a) was Einstein's flash of insight that elsewhere he called 'the happiest thought of my life'.[26] Part 3(b) is Einstein's ***equivalence principle***,[27] which he seems to have arrived at through 3(a). Although at first sight 3(a) and 3(b) appear to be related by interchanging the perspectives of the two observers involved (and indeed are often conflated), in fact they are quite different:

3(a) The modern way of phrasing this would be that in the frame of reference of an observer moving along a geodesic (i.e. a freely falling observer), *on the geodesic* the metric is the Minkowski metric and even its *first* derivatives vanish, so that also the Christoffel symbols vanish and hence the geometry of space-time as well as the motion of freely falling particles accompanying our observer are approximately flat.[28] Nonetheless, the gravitational field is not completely "transformed away" for the freely falling observer: the Riemann curvature tensor (which involves *second* derivatives of the metric) is nonzero even on the geodesic, and tidal forces (which are described by the Riemann tensor) are still there (cf. §5.1). Einstein understood this well before GR was established and hence 3(a) was *not* his equivalence principle; confusingly to many,[29] it was a heuristic towards it.

3(b) Version (b), which in turn became a crucial heuristic for Einstein in finding GR, is different from merely changing perspective in version (a). That would mean that instead of identifying with the freely falling observer, one identifies with Einstein sitting in his office, watching the man fall. Einstein then feels a pull downward *by a gravitational force*, which is what according to version (a) the freely falling man does *not* feel. However, Einstein's equivalence principle takes place in Minkowski space-time,[30] and states that a system undergoing *constant* acceleration may, as far as *all* laws of nature are concerned, equivalently considers itself at rest in a *homogeneous* gravitational field. Think of pushing the gas pedal in a car (preferably with blinded windows); the backward pull the driver feels from the acceleration is indistinguishable from a horizontal gravitational pull.

[25] See Proposition 7.2 and footnote 319. One must concede that the idea is clearer than its mathematical execution.

[26] 'Da kam mir der glücklichste Gedanke meines Lebens.' See Einstein (2002a), Doc. 31, page 265.

[27] Two leading scholarly papers on Einstein's equivalence principle are Norton (1985) and Lemkuhl (2019).

[28] See the end of §5.2 for the mathematical underpinning of this claim through Fermi normal coordinates.

[29] Perhaps starting with Pauli (1921, §51), who states that 'For the general case, [the equivalence principle] can be formulated in the following way: *For every infinitely small world region (. . .) there always exists a coordinate system (. . .) in which gravitation has no influence either on the motion of particles or any other physical processes*.'

[30] In 1907 Einstein had not yet internalized this concept and thought in terms of $3d$ "relative space", but in later formulations he talks about 'spacetime regions' in 'the limiting case of special relativity.' See Norton (1985), §2.

We would now say that version (a) is about *real* gravitational fields, whereas version (b) is about *fictitious* gravitational fields, whose properties are studied from their claimed equivalence with the effects of accelerated motion in Minkowski space-time, *as if* they were real. But apparently Einstein, whose views were often different from our modern ones, thought of them as real![31]

This can be made more precise by using comoving coordinates (with the accelerated observer), in which Christoffel symbols appear, of the same kind that locally describe "real" gravity (see also §1.10). Accordingly, Einstein's strategy, effective from 1912 onwards, was to infer properties of real gravity from known properties of accelerated motion, reinterpreted as the effects of fictitious gravity (as we see it, or as real gravity as Einstein saw it) as felt by an observer believing to be at rest (instead of accelerating). Its most important application, dating from 1912, is Einstein's crucial insight that gravity requires curved space-time, and hence should be based on differential geometry. This was the mathematical key to GR. With hindsight, the argument is quite straightforward: in the usual coordinates (t,x,y,z) the Minkowski line element is

$$ds^2 = -c^2dt^2 + dx^2 + dy^2 + dz^2, \tag{1.1}$$

but in arbitrary coordinates (seen by Einstein as describing an arbitrary reference frame) it is

$$ds^2 = g_{\mu\nu}(x)dx^\mu dx^\nu, \tag{1.2}$$

where $x \equiv (x^\mu) \equiv (x^0,x^1,x^2,x^3)$, and the $g_{\mu\nu}(x)$ are certain functions (jointly forming a Lorentzian metric at each point of space-time, as we would now say). According to version (b) of the equivalence principle, then, an observer who is accelerating with respect to the coordinates (t,x,y,z) may use comoving coordinates (x^μ) in which he is entitled to feel at rest in a gravitational field. At the same time, his line element is (1.2) rather than (1.1), and so he attributes the effects of this field to the functions $g_{\mu\nu}$. Einstein's leap of faith–one of the most successful in the history of science–was to generalize this argument from gravitational fields "caused" by acceleration–whatever their ontological status–to all gravitational fields, and hence claim that gravity is described by a metric tensor (or, as he later proposed, by its Christoffel symbols).

In actual fact, this insight came to Einstein in two steps, the first related to curved *time* and the second to curved *space*. First, in a constantly and linearly accelerated reference frame (t',x',y',z'), which his equivalence principle was always about, the line element (1.2) becomes

$$ds^2 = -c^2(x',y',z')(dt')^2 + (dx')^2 + (dy')^2 + (dz')^2, \tag{1.3}$$

so by version (b) of the equivalence principle gravity changes the metric in the time-like direction.[32] Second, consider a child moving on a merry-go-round, with a parent waiting outside. According to the latter, the child is accelerated inward, but according to the equivalence principle version (b) the child (who has studied *general* relativity) may claim to be at rest in a gravitational field that is radially outward directed and gives the centrifugal pull the child feels when it holds tight to the wooden horse. Now return to the parent, who knows *special* relativity (which without gravity is enough), and hence knows that moving objects contract in the direction of motion. This affects the length of measuring rods tangent to the circumference of the disc, but not those perpendicular to it (i.e. lying in the radial direction). Consequently, to the parent the ratio circumference/radius exceeds 2π. By the equivalence principle this is also true for the prodigy, who thereby concludes that gravity requires non-Euclidean geometry, at least spatially.[33]

[31] See Norton (1985) and Lemkuhl (2019). Accordingly, as part of his unholy alliance with Mach's principle (see footnote 24). Einstein attempted to find a material source of this induced gravitational field "at infinity".

[32] This leads to the prediction of gravitational deflection of light, which was *the* most famous early test of GR.

[33] See Stachel (1980). Einstein's argument is more complicated than it sounds, since the description of uniformly rotating solid discs in special relativity is tricky. However, it was just a heuristic and should be seen as such.

1.2 Riemannian geometry

Fortunately, as a student at the ETH Zürich, where he studied from 1896–1900 to become a mathematics teacher (*Fachlehrer in mathematischer Richtung*), Einstein had taken courses in differential geometry (*Infinitesimalgeometrie*) and geometric invariant theory (*Geometrische Theorie der Invarianten*) from a good mathematician called Carl Friedrich Geiser (1843–1934).[34] Though elementary, these courses were exactly what Einstein needed in 1912 to orient himself towards the right mathematical framework for GR (as he later acknowledged): apart from connections to the theory of functions, the differential geometry course for example discussed topics like coordinate systems, surfaces, line elements, (Gaussian) curvature, and geodesics.

Einstein's second stroke of luck was that in August 1912 he had moved from Prague to Zürich to become a professor at the ETH, where he renewed his friendship with his former ETH classmate Marcel Grossmann (1878–1936), who had been a professor of mathematics at the ETH since 1907. Grossmann was an expert in non-Euclidean geometry and introduced Einstein to the latest developments in this area.[35] Non-Euclidean geometry had started secretly with Carl Friedrich Gauss (1777–1855), one of the greatest mathematicians of all time, who however during his lifetime only published his technical work on lines and surfaces embedded in Euclidean space $\mathbb{R}^3$ that launched the field of differential geometry (Gauss, 1828).[36] This published work includes the description of curvature and the *Theorema Egregium* (i.e. 'remarkable theorem'), which states that what we now call the Gaussian curvature of a surface, though initially defined via its embedding in $\mathbb{R}^3$ (i.e. extrinsically), is in fact independent of the embedding and hence is an intrinsic property. See eq. (4.76) in §4.3. Thus a surface has both intrinsic and extrinsic curvature, a fact which–jumping ahead of Einstein–in one dimension higher (namely a three-dimensional space embedded in a four-dimensional space-time) will play a central role in the initial-value problem of GR (as well as in its closely related Hamiltonian formulation). See chapter 8.

The work of Gauss was taken a decisive step further by his brilliant–perhaps even greater–student Bernhard Riemann (1826–1866). In his extraordinarily visionary Habilitation lecture on June 10, 1854, Riemann simply left out the ambient Euclidean space and also worked in arbitrary dimension.[37] This lecture starts in the following provocative way:[38]

[34] Einstein in fact rarely went to lectures and prepared himself for exams using notes taken by Grossmann.

[35] See Sauer (2014) for a survey of Grossmann's interaction with Einstein and his contributions to GR.

[36] A very good introduction to this work is Volume 2 of Spivak (1999), chapter 3, which includes a translation.

[37] Also here Volume 2 of Spivak (1999), chapter 4, is an excellent introduction. Riemann's lecture from 1854 was given for a non-mathematical audience including philosophers (but also the aging Gauss) and so it contains almost no equations. What we now call the Riemann tensor was first given in an initially unpublished prize essay on heat conduction, known among historians as the *Commentatio*, which first appeared in Riemann (1876), pp. 370–383. In this essay, Riemann models heat flow using a three-dimensional metric, whose local flatness (which turned out to be physically interesting) he relates to the vanishing of the curvature tensor, cf. Theorem 4.1 in the present book. See Farwell & Knee (1990) and Darrigol (2014). The transition from Gauss to Riemann is described in detail by Reich (1973), and is embedded in the general history of ninetheenth century geometry by Gray (2007).

[38] 'Bekanntlich setzt die Geometrie sowohl den Begriff des Raumes, als die ersten Grundbegriffe für die Constructionen in Raume als etwas Gegebenes voraus. Sie giebt von ihnen nur Nominaldefinitionen, während die wesentlichen Bestimmungen in Form von Axiomen auftreten. Das Verhältniss dieser Voraussetzungen bleibt dabei in Dunkeln; man sieht weder ein, ob und in wie weit ihre Verbindung nothwendig, noch a priori, ob sie möglich ist. Diese Dunkelheit wurde auch von Euklid bis auf Legendre, um den berühmtesten neueren Bearbeiter der Geometrie zu nennen, weder von den Mathematikern, noch von den Philosophen, welche sich damit beschäftigten, gehoben. Es hatte dies seinen Grund wohl darin, dass der allgemeine Begriff mehrfach ausgedehnter Grössen, unter welchem die Raumgrössen enthalten sind, ganz unbearbeitet blieb. Ich habe mir daher zunächst die Aufgabe gestellt, den Begriff einer mehrfach ausgedehnten Grösse aus allgemeinen Grössenbegriffen zu construiren. Es wird daraus hervorgehen, dass eine mehrfach ausgedehnte Grösse verschiedener Massverhältnisse fähig ist und der Raum also

> It is known that geometry assumes, as things given, both the notion of space and the first principles of constructions in space. She gives definitions of them which are merely nominal, while the true determinations appear in the form of axioms. The relation of these assumptions remains consequently in darkness; we neither perceive whether and how far their connection is necessary, nor a priori, whether it is possible.
>
> From Euclid to Legendre (to name the most famous of modern reforming geometers) this darkness was cleared up neither by mathematicians nor by such philosophers as concerned themselves with it. The reason of this is doubtless that the general notion of multiple extended magnitudes (in which space-magnitudes are included) remained entirely unworked. I have in the first place, therefore, set myself the task of constructing the notion of a multiply extended magnitude out of general notions of magnitude. It will follow from this that a multiply extended magnitude is capable of different measure-relations, and consequently that space is only a particular case of a triply extended magnitude. (Riemann, 1854)

As Gray (2007, p. 193) put it, 'Euclid's postulates are completely subverted: no longer can they be regarded as unproblematically true assumptions about physical space.' Even in two dimensions, the (hyperbolic) non-Euclidean geometries discovered two decades earlier by Bolyai and Lobachevskii were far from the only possibilities (although, as Riemann mentioned, they do have special symmetry properties).[39] Riemann's main ideas, hardly formalized by him however, were firstly that of a *manifold* (described by him as an 'n-fach ausgedehnte Grösse', and later even as a 'Mannigfaltigkeit'),[40] and secondly that of a *metric* (defined on a manifold), which he made responsible for derived notions like distance, angles, geodesics, and curvature, and as such identified as the basis of geometry. Applied to space-time rather than space,[41] this turned out to be exactly what was needed for GR. This gives the second great and remarkable example of a piece of mathematics that was initially developed for purely intrinsic (i.e. mathematical) reasons but later turned out to provide the right language for some profound new physical theory.[42]

---

nur einen besonderen Fall einer dreifach ausgedehnten Grösse bildet.' Translation by W.K. Clifford.

[39] Riemann did not name Bolyai and Lobachevskii and probably did not know their work. Yet one of the few formulas in his lecture gives the metric of hyperbolic space as an example of a space with constant curvature.

[40] The historical development of the concept of a manifold is described by Scholz (1980, 1999). The word 'Mannigfaltigkeit' had been used by Gauss in lectures, but always in the context of subspaces of (what we now call) $\mathbb{R}^n$, and it really seems to have been Riemann who conceived the general notion, including hints towards global structure described by overlapping charts (one may even argue that his habilitation lecture foreshadowed both set theory and topology; for example, he explicitly left room for discrete as opposed to continuous structures, and in his earlier PhD thesis from 1851 Riemann had even talked about infinite-dimensional spaces of functions). However, as Scholz notes (p. 30), 'The reception and assimilation of Riemann's concept of a manifold to the mathematics of the 19th century was slow and inhibited by severe conceptual problems'. As we shall see, Ricci (and Levi-Civita) even turned the clock back by omitting any reference to global structure and basing their tensor calculus entirely on the use of coordinates without a specified domain (which was often implicitly taken to be $\mathbb{R}^n$). Nonetheless, in a development in which topology, geometry, and function theory can hardly be separated, through the work of Beltrami, Helmholtz, Klein, Möbius, Jordan, Schäfli, Betti, Poincaré, Brouwer, Hausdorff and others, the modern notion of a manifold finally arose. In dimension two, Hilbert (1902b) sketched the modern definition in terms of open neighbourhoods, charts, and coordinate changes (where he also had to define the fundamental notions of topology, a subject that at the time had by no means been brought into final form). This was subsequently formalized by Weyl (1913) in dimension two, and then by Veblen & Whitehead (1932) and Whitney (1936) in general.

[41] Even this twist we owe to a mathematician, namely Minkowski, cf. Corry (2004).

[42] The first great example is the application of the conic sections of the ancient Greeks (first described in the fourth century B.C. in the context of problems in Euclidean geometry) to motion in a gravitational field, starting with Galilei's parabolic motion of projectiles on earth and culminating in Newton's derivation of Kepler's laws describing the elliptic motion of planets in *Principia*–one of the highlights in the history of science, on a par with the discovery of GR. The second great example, then, is GR. The third is functional analysis, which developed out of abstract

1.3 Absolute differential calculus and general covariance

To this remarkable success story one should add the computational device that made Riemannian geometry workable for Einstein, namely the "absolute differential calculus" developed by Gregorio Ricci-Curbastro (1853–1925), also simply called Ricci.[43] This calculus was written down in final form in 1901 in a joint paper by Ricci and his student Tulio Levi-Civita (1873–1941), who also became interested in GR (including a personal friendship with Einstein).[44]

Of historical interest only now, this paper is very instructive as a portrait of the mathematical world which Einstein inhabited in the 1910s. The absolute differential calculus (or tensor calculus) uses formal real variables $x_1, x_2, \ldots x_n$, but any kind of geometric perspective is absent: the calculus is a completely formal mix of algebra and analysis. Even the abstract framework of (multi)linear algebra is lacking (multilinear maps are written in terms of their components relative to a basis), so that everything is written down in terms of indices and tensors are defined (as they still are in some modern physics books) by their behaviour under coordinate transformations. The main achievement of the absolute differential calculus is the introduction of the covariant derivative on arbitrary tensors, along with all the rules for working with it.[45] This gives the Riemann tensor, the Ricci tensor, the Ricci scalar, and many other similar constructions, studied from the point of view of invariant theory (as opposed to geometry).

The next stage in Einstein's path to GR only makes sense if we understand Einstein's conflation of general covariance with a relativity principle.[46] It is crucial to realize that for us, coordinate systems are arbitrary, physically dead *labelings* of points in space-time. But for Einstein, coordinates were alive as physical *frames of reference*, in the sense that the system $(x^0, \vec{x})$ really describes the world line $(x^0(t), \vec{x}(t)) = (t, \vec{x})$ of an observer who is spatially at rest at $\vec{x}$, but moves in time t, including the stipulation that events at $(t, \vec{x})$ and $(t, \vec{y})$ are simultaneous,

---

nineteenth century analysis and turned out to be exactly the right mathematical language for quantum mechanics (e.g. Landsman, 2019). This phenomenon is still not well understood. Hilbert and his circle, who played a key role in both the second and the third example (i.e. GR and quantum theory), invoked what they called a "pre-established harmony between physical nature and mathematical mind" (Corry, 2004), but this seems a sledge-hammer argument that explains nothing. Note that the issue is *not* what Wigner (1960) famously called the 'unreasonable effectiveness of mathematics in the natural sciences', or, in other words, the 'appropriateness of the language of mathematics for the formulation of the laws of physics', which, he added lyrically but misleadingly, 'we neither understand nor deserve'. Without in any way lessening our admiration for Newton's genius, we perfectly well understand the applicability of the calculus to classical mechanics, since Newton purposely developed those in close interaction with each other. Our point is that the conic sections were already there, waiting for him. The miracle, if there is one, is the applicability of mathematical concepts that were invented purely for their own sake to physical theories like GR and quantum mechanics, which postdate these inventions with no apparent link or common cause.

[43] See Reich (1994) for the relevant mathematical history in depth and Goodstein (2018) for (light) biography.

[44] Levi-Civita later wrote a textbook on the absolute differential calculus including its application to GR (Levi-Civita, 1926; Italian original from 1923), in which he uses the concept of parallel transport he had invented himself in the wake of GR. This makes the 1923 book slightly more geometric than the 1901 paper, but most of the comments in the main text about the 1901 paper also apply to Levi-Civita's book. Almost simultaneously, the Dutch mathematician Jan Arnoldus Schouten (1883–1971) published his book Schouten (1924), dedicated to Ricci, which is similar, to Levi-Civita's book except that it only mentions Einstein in a footnote as someone who applied the theory of linear connections to physics ('Physikalische Anwendungen gaben Weyl, Eddington, und Einstein'), and leaves it at that as far as GR is concerned. Schouten founded a Dutch school in tensor calculus that involved e.g. Dirk-Jan Struik (1894–2000), who later became a well-known (Marxist) historian of mathematics, and Max Euwe (1901–1981), who was originally a mathematics high-school teacher but is better known from his career in chess, in which he was world champion from 1935–1937. He later became one of the first Dutch computer scientists.

[45] The Lie derivative is still absent from the tensor calculus; it was introduced in 1931 by the Polish mathematician Władysław Ślebodziński (1884–1972), who survived Auschwitz and two other concentration camps.

[46] Relevant literature, none of which we literally follow, includes Norton (1989, 1993) and Janssen (2012, 2014).

at least for the observer moving along the given world line.[47] In order to even talk about, for example, the speed of light, the coordinate t must then be given the physical meaning of time, and the coordinate difference $|\vec{x}-\vec{y}|$ the physical meaning of distance. For Einstein, then, inertial frames are described by distinguished coordinate systems, and coordinate transformations correspond to changes in frames of reference, which may or may not preserve inertial frames. In special relativity inertial frames correspond to geodesics,[48] and hence to do justice to the two ingredients of the principle of special relativity (i.e. relativity of uniform motion and constancy of the speed of light) it seems natural to define symmetries as transformations (i.e. diffeomorphisms) of space-time that map geodesics into geodesics and preserve the speed of light. This precisely gives the Poincaré transformations (which are Lorentz transformations combined with constant translations in space-time), which in turn coincide with the isometries of the Minkowski metric.[49]

Einstein's reasoning then seems to have been as follows. The special principle of relativity states that the laws of physics (including constancy of the speed of light but excluding gravity) are the same in each inertial frame (and in no others). Hence the *special* principle of relativity is equivalent to the invariance of the laws of physics under certain *special* coordinate transformations, namely Poincaré transformations. Therefore, the *general* principle of relativity (which Einstein was after because he liked Mach's principle and in special relativity disliked the presence of special coordinate systems–which he identified with inertial reference frames) should consist of the invariance of the laws of physics under *general* coordinate transformations.[50]

By a stroke of fortune Ricci's absolute differential calculus gave Einstein partial differential equations for physics that were invariant under *general* coordinate transformations; this was even what Ricci meant by "absolute". And this, in Einstein's view, made all physical frames of reference equivalent and gave him the mathematical machinery for his "general principle of relativity". The equivalence principle then implied that general relativity is only possible in the presence of gravity, indeed *is* a theory of gravity, which is then automatically generally covariant.

The requirement of general covariance was one of the keys for Einstein in finding his field equations during the years 1913–1915, though not without a distraction in the form of the Hole Argument, as we shall see shortly. Later in his life Einstein increasingly came to believe that mathematics (and in particular the idea of general covariance) had been *the* key to his success, which (not even mentioning his own physical insights) already in 1915 he had described as a 'real triumph of the general method of the differential calculus developed by Gauss, Riemann, Christoffel, Ricci, and Levi-Civita.' His most blatant statement in this direction is probably that:

[47] Hence a reference frame should perhaps be taken to be a congruence of geodesics, rather than a single one.

[48] It was Einstein who reintroduced the geometric concept of a geodesic in this context–a crucial move towards the current reconciliation of the tensor calculus with differential geometry–but formulated in terms of coordinates.

[49] The diffeomorphisms of $\mathbb{R}^4$ that merely preserve the geodesics of the Minkowski metric just have to preserve straight lines and hence correspond to affine maps, i.e., linear transformation plus translation. This is also true in Euclidean space, where affine transformations preserve straight lines but only isometries also preserve distances. In the Euclidean case the linear part of an isometry must be a rotation or a reflection, whereas in the Minkowski case it must be a Lorentz transformation (which notion by definition includes spatial and temporal reflections). See also Kobayashi & Nomizu (1963), chapter VI, for the notion of affine transformations of manifolds with an affine connection, such as the Levi-Civita connection, and, in that case, their relationship to isometries.

[50] We return to this issue in §1.10. For now, we just mention that Einstein's argument is widely regarded as suspicious and that the correct generalization of his reasoning about special relativity would be to say that symmetries *of a specific space-time* (including a metric) are *isometries* (which in particular map geodesics to geodesics), whereas the symmetries *of* GR *as a theory* are *diffeomorphisms* (or, for that matter, general coordinate transformations). Since any kind of relativity of motion should refer to some specific space-time, it would be a category mistake to infer it from invariance properties of the theory as a whole. If anything, motion is relative only with respect to (non-trivial) isometries of a fixed space-time (if these exist), which preserve geodesics and Lorentzian distances.

> The gravitational equations could *only* be found by a purely formal principle (general covariance), that is, by trusting in the largest imaginable logical simplicity of the natural laws.[51] (Einstein to De Broglie, 1954)

In his Herbert Spencer Lecture in Oxford, 1933, he mused:

> Newton (...) still believed that the basic concepts and laws of his system could be derived from experience (...). It was the general Theory of Relativity which showed in a convincing way the incorrectness of this view. For this theory revealed that it was possible for us, using basic principles very far removed from those of Newton, to do justice to the entire range of the data of experience in a manner even more complete and satisfactory than was possible with Newton's principles. But quite apart from the question of comparative merits, the fictitious character of the principles is made quite obvious by the fact that it is possible to exhibit two essentially different bases, each of which in its consequences leads to a large measure of agreement with experience. This indicates that any attempt logically to derive the basic concepts and laws of mechanics from the ultimate data of experience is doomed to failure. If then it is the case that the axiomatic basis of theoretical physics cannot be an inference from experience, but must be free invention, have we any right to hope that we shall find the correct way? Still more–does this correct approach exist at all, save in our imagination? Have we any right to hope that experience will guide us aright, when there are theories (like classical mechanics) which agree with experience to a very great extent, even without comprehending the subject in its depths? To this I answer with complete assurance, that in my opinion there is the correct path and, moreover, that it is in our power to find it. Our experience up to date justifies us in feeling sure that in Nature is actualized the ideal of mathematical simplicity. It is my conviction that pure mathematical construction enables us to discover the concepts and the laws connecting them which give us the key to the understanding of the phenomena of Nature. Experience can of course guide us in our choice of serviceable mathematical concepts; it cannot possibly be the source from which they are derived; experience of course remains the sole criterion of the serviceability of a mathematical construction for physics, but the truly creative principle resides in mathematics. In a certain sense, therefore, I hold it to be true that pure thought is competent to comprehend the real, as the ancients dreamed. (Einstein, 1934, pp. 166–167)

Similarly, Hilbert, to whose role in the development of GR we will return in §1.7, saw Einstein's theory as the final demise of the idea that physical theories should be based on experience:

> In Einstein's theory we now have a conistent field theory before us; the second stage in the development of physics has thereby been reached. What happens is not merely switching off the senses, as is the case with mechanics, but rather the complete elimination of anthropomorphism. The conceptual structures have completely emancipated themselves from the usual sense impressions, and it is precisely by getting rid of these that objectivity in the understanding of the laws of nature as well as the unity and clarity of the theoretical system are achieved. In this regard, I would like to regard the general principle of relativity as the highest triumph of the mind over the world of appearances.[52] (Hilbert, 1919/1920)

[51] 'Die Gravitationsgleichungen waren *nur* auffindbar auf Grund eines rein formalen Prinzips (allgemeine Kovarianz), d.h. auf Grund des Vertrauens auf die denkbar grösste logische Einfachheit der Naturgesetze.' Quoted in van Dongen (2010), pp. 2–3, whose book is a major analysis of the issue at hand. See also van Dongen (2017).

[52] 'In dieser Einsteinschen Theorie haben wir nun eine konsequente Feldtheorie vor uns; die zweite Stufe in der

1.4 Towards the gravitational field equations: *Entwurf Theorie*

However, historical reconstruction has shown that the truth may have been quite different. Considerable evidence shows that Einstein did find his equations by the mathematical requirement of general covariance, *but combined with various physical requirements he always had in mind*,[53] notably the necessity of the correct Newtonian limit as well as of energy-momentum conservation.

Specifically, after he had realized that Newton's gravitational force (or rather its scalar potential φ) should be replaced by the 10 components of the metric tensor $g_{\mu\nu}$, and through Grossmann had familiarized himself with the necessary mathematics, from the autumn of 1912 onwards Einstein actively tried to involve the metric $g_{\mu\nu}$ in generalizing Poisson's equation

$$\Delta\varphi = -4\pi G\rho, \tag{1.4}$$

where G is Newton's gravitational constant and ρ is the matter density. He aimed at the structure

$$Q_{\mu\nu} = \kappa T_{\mu\nu}, \tag{1.5}$$

where $T_{\mu\nu}$ is the energy-momentum tensor that had already been introduced in special relativity by von Laue and had been generalized to curved space-time by Kottler, κ is some constant (which later became $\kappa = 8\pi G$), and $Q_{\mu\nu}$ is some tensor to be constructed from the metric using the absolute differential calculus.[54] It took Einstein three years to get to the correct expression

$$Q_{\mu\nu} = R_{\mu\nu} - \tfrac{1}{2}g_{\mu\nu}R, \tag{1.6}$$

during which he 'dedicated himself to the problem of gravitation with superhuman effort'.[55]

In retrospect he was almost there right from the start, since after Grossmann had pointed out the Riemann tensor to him in the autumn of 1912 Einstein at once tried the associated Ricci tensor $Q_{\mu\nu} = R_{\mu\nu}$, but this turned out to give the wrong Newtonian limit–or so he thought; in fact, the problem did not lie in the omission of the $-\frac{1}{2}g_{\mu\nu}R$ term, but with the coordinates he used, as well as with a misconception that would trouble Einstein for years to come, namely that in the Newtonian limit (and in suitable coordinates) the metric should take the diagonal form (1.3), where the variable speed of light $c(x',y',z')$ takes care of "everything". In other words, time is curved but space remains Euclidean. As the Schwarzschild solution shows, this is wrong,

In his search for the gravitational field equations Einstein was led by a powerful formal analogy with electrodynamics (whose four-dimensional formulation due to Minkowski he had initially been slow to endorse), whose ("specially" covariant) field equations take the form

$$\partial_\rho F^{\mu\rho} = kJ^\mu, \tag{1.7}$$

Entwicklung der Physik ist damit erreicht. Nicht bloß eine Ausschaltung der Sinne, wie bei der Mechanistik, findet hier statt, sondern eine gänzliche Beseitigung des Anthropomorphismus. Die Begriffsbildungen haben sich ganz und gar von dem anschaulig Geläufigen emanzipiert; und gerade dadurch, daß man sich von der Anschauung losmacht, wird die Objektivität in der Auffassung der Naturgesetze sowie die Einheit und Übersichtlichkeit des theoretischen Systems erreicht. In dieser Hinsicht möchte ich das allgemeine Relativitätsprinzip als den höchsten Triumph des Geistes über die Erscheinungswelt ansehen.' (Hilbert, 1992, p. 51).

[53] See Janssen (2014). Ironically, Einstein started his Herbert Spencer Lecture with the following warning: 'If you wish to learn from the theoretical physicist anything about the methods which he uses, I would give you the following piece of advice: Don't listen to his words, examine his achievements.' See also van Dongen (2010).

[54] The reconstruction of Einstein's ideas during 1912–1913 is largely based on the *Zürich Notebook*, which Einstein used from August 1912 to May 1913. A transcription may be found in Einstein (1996a) and a fascimile with transcription and commentary is in Renn (2007), Volume 1, pp. 313–487. The original is kept in Jerusalem.

[55] 'geradezu übermenschlichen Anstrengungen, mit denen ich mich dem Gravitationsproblem gewidmet habe', as he wrote on May 28, 1913, to his friend Paul Ehrenfest, quote taken from Fölsing (1993, p. 357).

where F is the electromagnetic field strength tensor, $J = (J^0, \vec{J})$ is the electric (charge density, current), and k is a constant. This suggested to Einstein that (1.5) should also be thought of as

$$\partial_\rho H^\rho_{\mu\nu} = \kappa(T_{\mu\nu} + t_{\mu\nu}), \tag{1.8}$$

where the object $H^\rho_{\mu\nu}$, constructed from the metric, represents the gravitational field, and $t_{\mu\nu}$ is the energy-momentum tensor of the gravitational field itself (whereas $T_{\mu\nu}$ is the energy-momentum tensor of the matter in the universe). From the point of view of (1.5), an obvious first guess for the left-hand side, which Einstein indeed wrote down, is (up to a constant factor)

$$Q_{\mu\nu} = g^{\rho\sigma}\frac{\partial^2 g_{\mu\nu}}{\partial x^\rho \partial x^\sigma}, \tag{1.9}$$

where $(g^{\rho\sigma})$ is the inverse matrix to $(g_{\rho\sigma})$ as usual, but in the context of (1.8) he started from

$$H^\rho_{\mu\nu} = -\tfrac{1}{2}g^{\rho\sigma}\partial_\mu g_{\sigma\nu}, \tag{1.10}$$

which he later described as a 'fateful prejudice'. It was only in November 1915 that he realised that the choice

$$H^\rho_{\mu\nu} = -\Gamma^\rho_{\mu\nu}, \tag{1.11}$$

where $\Gamma^\rho_{\mu\nu}$ are the Christoffel symbols he knew well from the absolute differential calculus, i.e.,

$$\Gamma^\rho_{\mu\nu} = \tfrac{1}{2}g^{\rho\sigma}(\partial_\nu g_{\sigma\mu} + \partial_\mu g_{\sigma\nu} - \partial_\sigma g_{\mu\nu}), \tag{1.12}$$

gave him the best of both worlds (though not yet quite the correct field equations, see below). One reason for (1.10) may have been that if he adapted the Lagrangian of electrodynamics, i.e.

$$\mathscr{L} = -\tfrac{1}{4}F_{\mu\nu}F^{\mu\nu} = -\tfrac{1}{4}\eta^{\mu\rho}\eta^{\nu\sigma}F_{\mu\nu}F_{\rho\sigma}, \tag{1.13}$$

to the gravitational case by postulating

$$\mathscr{L} = -g^{\nu\sigma}H^\rho_{\mu\nu}H^\mu_{\rho\sigma}, \tag{1.14}$$

then the choice (1.10) led to what historians call the ***Entwurf Theorie*** (Einstein & Grossmann, 1913). Although the field equations of this theory (whose tedious explicit form we omit), derived from (1.14) by the variational calculus, are not generally covariant (like the Lagrangian itself), Einstein nonetheless felt they were correct, since they gave both the Newtonian limit and energy-momentum conservation, albeit only in certain preferred coordinate systems. He wrote:

> The labor is finally ready, after endless trouble and vexing doubts.[56]

This brought Einstein in a very interesting psychological situation: since he believed his theory was ready despite the egregious shortcoming of not being generally covariant (and hence, as he thought, not satisfying the "general principle of relativity", and hence, via his virtual identification of all these things, violating the equivalence principle), he started looking for arguments *against* general covariance (and, by implication, almost all his other holy principles)![57]

[56] 'Die Arbeit ist nach unendlicher Mühe und quälenden Zweifeln nun endlich fertig geworden.' Quote from an undated letter to Ernst Mach, probably written during the summer of 1913, taken from Fölsing (1993).

[57] See Janssen (2014) and Norton (2018) for fascinating reflections on this remarkable aspect of Einstein's mind.

1.5 The Hole Argument

Whatever its origin, one such argument has been of lasting value, namely the ***Hole Argument***(*Lochbetrachtung*).[58] Before turning to Einstein's rendition of this argument, it may be helpful, though anachronistic, to first give Hilbert's (1917) reformulation of Einstein's Hole Argument, which was also the first attempt to discuss the Cauchy problem for the Einstein equations (see §1.9). Adding insult to historical injury, we interpret general covariance of GR as diffeomorphism invariance, in the sense that if a metric g solves the vacuum Einstein equations

$$R_{\mu\nu} = 0; \qquad \text{i.e.} \qquad \mathrm{Ric}(g) = 0, \tag{1.15}$$

then also $\psi^* g$ solves these equations, for any diffeomorphism ψ. This is because

$$\mathrm{Ric}(\psi^* g) = \psi^* \mathrm{Ric}(g), \tag{1.16}$$

cf. §2.5.4. That is, the Ricci tensor for the transformed metric $\psi^* g$ is the transform (pullback) of the Ricci tensor for the original metric g (this property is also the root of the Bianchi identities). If we now take initial data on some three-dimensional spacelike hypersurface Σ (it will be explained in §7.6 what this exactly means), and find a diffeomorphism ψ that is equal to the identity in some neighbourhood of Σ but is nontrivial elsewhere, and g solves the Einstein equations with the given initial data, then so does $\psi^* g$, for the same initial data! Hence a generally covariant theory cannot be deterministic in the sense of being a system of partial differential equations with a well-posed initial value (Cauchy) problem with a unique solution (at least for short times).

The name "Hole Argument" comes from Einstein's original version. Turning the above story inside out, he takes the region where ψ (which in his case is a coordinate transformation) is nontrivial to be a (four-dimensional) "hole" in space-time, and notes that boundary conditions outside the hole do not determine the metric within it.[59] In coordinates the argument reads as follows. If $g_{\mu\nu}(x)$ solves some generally covariant equations in a coordinate system (x), then so does $g'_{\mu\nu}(x')$, i.e. the same metric expressed in a new coordinate system (x'), constructed from $g_{\mu\nu}(x)$ by the usual transformation rules for tensors. *This is the same metric.* Einstein's point is that $g'_{\mu\nu}(x)$ also solves the equations, even though (barring isometries), *it is a different metric*. Therefore, Einstein (correctly) concluded, the gravitational field is not uniquely fixed.

Strangely enough, during 1915, when he returned to general covariance, Einstein conveniently forgot to mention his Hole Argument, returning to it only in his review Einstein (1916a), preceded by discussions in private correspondence from December 1915 onward. Based on the so-called *point-coincidence argument*, which claims that 'nothing is physically real but the totality of space-time point coincidences',[60] he argued that (M, g) and $(M, \psi^* g)$ represent the same physical situation, i.e., in modern terminology, that diffeomorphisms are gauge symmetries.[61] For the moment we leave it at that, and return to development of GR during 1913–1915.

[58] See Stachel (2014), Norton (2019), and Pooley (2020) for surveys, and e.g. Weatherall (2018) for further analysis. Another argument that at the time convinced Einstein that the *Entwurf Theorie* was correct (and hence general covariance was untenable) was that energy-momentum conservation was only possible in specific coordinate systems, namely precisely in those where the *Entwurf* field equations were supposed to be valid.

[59] This looks unnatural compared to Hilbert's formulation, but as Stachel (2014) remarks, Einstein was probably once again inspired by Mach's principle, where "the fixed stars at infinity" determine the local intertia of matter.

[60] Quoted in Stachel (2014) and elsewhere from a letter to Besso, 3 January 1916, translation by Stachel.

[61] Though this answers the Hole Argument, it is not actually true in asymptotically flat space-times, where diffeomorphisms induce physically nontrivial transformations at infinity. Furthermore, interpreting diffeomorphisms as gauge symmetries leads to the problem of time, which we will return to in later in this book, cf. §8.11.

1.6 Finding the gravitational field equations: November 1915

During these years, Einstein became increasingly dissatisfied with his *Entwurf Theorie*:[62]

> I recognized that the field equations for gravitation I had so far were totally untenable.[63]

It is remarkable how quickly Einstein then collected himself, since in the dramatic month of November 1915 he wrote four brief papers converging to the final answer (1.6), although, as if they were a compressed history of the preceding years, some of these contained new mistakes.[64] One reason for Einstein's hurry in putting his thoughts into print was a competition with Hilbert (or at least that is what Einstein felt); we will return to Hilbert's role in the history of GR shortly.

The first paper (Einstein, 1915a, dated November 4) still failed to achieve general covariance, but at least Einstein states the intention to restore it (in a remarkably personal passage):

> On these grounds I completely lost confidence in the field equations I had established and searched for a way to restrict the possibilities in a natural manner. Thus I got back to the requirement of more generally covariant field equations, which I had left only with a heavy heart when I worked together with my friend Grossmann. In fact we had already then come very close to the solution of the problem given in what follows.[65] (Einstein, 1915a, p. 778)

Einstein recognized that (1.11) rather than (1.10), which had led to the *Entwurf* equations, was the correct choice.[66] Putting (1.11) in the Lagrangian (1.14) then leads to the field equations

$$\tilde{R}_{\mu\nu} = \kappa T_{\mu\nu}, \tag{1.17}$$

where the non-covariant expression $\tilde{R}_{\mu\nu} = \partial_\rho \Gamma^\rho_{\mu\nu} - \Gamma^\rho_{\nu\sigma}\Gamma^\sigma_{\rho\mu}$ is "half" of the full Ricci tensor

$$R_{\mu\nu} = \partial_\rho \Gamma^\rho_{\mu\nu} - \partial_\nu \Gamma^\rho_{\mu\rho} + \Gamma^\rho_{\rho\sigma}\Gamma^\sigma_{\nu\mu} - \Gamma^\rho_{\nu\sigma}\Gamma^\sigma_{\rho\mu}. \tag{1.18}$$

[62] Einstein's rejection of the *Entwurf Theorie* is a story by itself, but briefly: (i) it did not satisfy Mach's principle as Einstein saw it (he insisted that a uniformly rotating empty Minkowski space-time should be a solution to the *Entwurf* equations, which it wasn't–a calculation Einstein apparently did over and over again with different results each time); (ii) there were problems with its Lagrangian formulation; (iii) Einstein's earlier arguments that the theory was unique given the correct Newtonian limit and energy-momentum conservation turned out to be flawed; and (iv) it got the perihelion shift of Mercury wrong (off by a factor 2.4), though Einstein's somewhat cynical reactions to such discrepancies between theory and experiment are too well known to be repeated here.

[63] 'Ich erkannte nämlich, dass meine bisherigen Feldgleichungen der Gravitation gänzlich haltlos waren!' Letter to Sommerfeld, 28 November 1915 (Einstein, 1999, Doc. 153). This insight refers to October 1915.

[64] Each of these papers was based on a talk Einstein gave at the Prussian Academy of Sciences on the day the paper is dated (in particular, he also presented his final field equations on November 25th). See Simon (2005). For those who wish to look at the original papers it is worth mentioning that Einstein denotes the Ricci tensor by G_{ik} instead of the current $R_{\mu\nu}$ (and today's $G_{\mu\nu}$ is the Einstein tensor $R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}R$), whilst his R_{ik} is minus our $\tilde{R}_{\mu\nu}$.

[65] 'Aus diesen Gründen verlor ich das Vertrauen zu den von mir aufgestellten Feldgleichungen vollständig und suchte nach einem Wege, der die Möglichkeiten in einer natürlichen Weise einschränkte. So gelangte ich zu der Forderung einer allgemeineren Kovarianz der Feldgleichungen zurück, von der ich vor drei Jahren, als ich zusammen mit meinem Freunde Grossmann arbeitete, nur mit schwerem Herzen abgegangen war. In der Tat waren wir damals der im nachfolgenden gegebenen Lösung des Problems bereits ganz nahe gekommen.'

[66] Compared to the Einstein–Hilbert Lagrangian $\mathscr{L}_{EH} = \sqrt{-g}R$, the Lagrangian $\mathscr{L}_{\text{Nov}\,4} = -g^{\nu\sigma}\Gamma^\rho_{\mu\nu}\Gamma^\mu_{\rho\sigma}$ used in Einstein (1915a), cf. (1.11) and (1.14), assuming $g = -1$, is not so far off. The first two terms in (1.18), which are absent in $\mathscr{L}_{\text{Nov}\,4}$, merely bring a divergence and hence do not contribute to the equations of motion; this is the reason why the Einstein equations are second-order, although $\mathscr{L}_{EH}$ contain second-order derivatives of the metric and hence *a priori* one would expect fourth-order equations. Furthermore, the third term in (1.18) vanishes if $g = -1$, cf. (1.19), so that all that survives of $\mathscr{L}_{EH}$ is precisely $\mathscr{L}_{\text{Nov}\,4}$. The reason the equations (1.17) miss $-\frac{1}{2}g_{\mu\nu}R$ is that this term arises from a variation of $\sqrt{-g}$ in $\mathscr{L}_{EH}$, which is missing in $\mathscr{L}_{\text{Nov}\,4}$ because it has been put equal to 1.

In *unimodular coordinates*, in which $g \equiv \det(g) = -1$, we actually have $\tilde{R}_{\mu\nu} = R_{\mu\nu}$, since

$$\partial_\mu \sqrt{-g} = \sqrt{-g}\,\Gamma^\rho_{\mu\rho}. \tag{1.19}$$

However, it seems that Einstein recognized this fact only after he had submitted his first November paper. For one had to wait for the second one (Einstein, 1915b) for him to say the following:

> This tensor $R_{\mu\nu}$ is the only tensor that is available for the formulation of generally covariant gravitational equations. If we now agree that the field equations of gravitation should be
>
> $$R_{\mu\nu} = \kappa T_{\mu\nu}, \tag{1.20}$$
>
> then we have gained generally covariant field equations.[67]

He then justifies (1.20) by the fact that in unimodular coordinates it coincides with his earlier (1.17), but notes a serious problem: combining (1.17) with the unimodularity condition

$$\det(g) = -1 \tag{1.21}$$

enforces $T^\mu_\mu = 0$. At this point, under less duress he would undoubtedly have seen that the problem is solved by adding $-\frac{1}{2}g_{\mu\nu}R$ to the left-hand side (or, equivalently, $-\frac{1}{2}g_{\mu\nu}T$ to the right-hand side) of (1.20). But this simple solution took him another week to arrive at.[68] Instead, he apparently felt compelled to save both his equations (1.17), which were the ones he really believed in, *and* general covariance. This combination required the unimodularity condition, and hence tracelessness of the energy-momentum tensor, which therefore had to be justified one way or the other. Such justification was available in the form of the *electromagnetic world hypothesis*, which went back to Gustav Mie (1868–1957), and also haunted Hilbert (as we shall see). In Einstein's case it was rather short-lived, since his only reason for believing in it was to obtain a traceless energy-momentum tensor. During the next week he saw his reasoning collapse once again, for he noted that the unimodularity condition was incompatible with what he still thought was the Newtonian limit of his theory, namely (1.3). But in Einstein (1915c) he redid the computation of the perihelion shift of Mercury, which he had first done with Besso in June 1913 using his *Entwurf Theorie*,[69] from his new equations (1.17), and assuming (1.21):

> Imagine my joy when I found that the equations correctly have the perihelion shift of Mercury (...) I was speechless from excitement for several days.[70]

The computation also opened Einstein's eyes to the incorrectness of (1.3) in the Newtonian limit and at last gave him the correct picture of it. Having rescued the condition (1.21), all that was left was to remove its undesired consequence $T \equiv T^\mu_\mu = 0$.[71] Thus Einstein (1915d) finally wrote

$$\boxed{R_{\mu\nu} = \kappa\left(T_{\mu\nu} - \tfrac{1}{2}g_{\mu\nu}T\right)}. \tag{1.22}$$

This was the end of his magnificent search for generally covariant gravitational field equations.

---

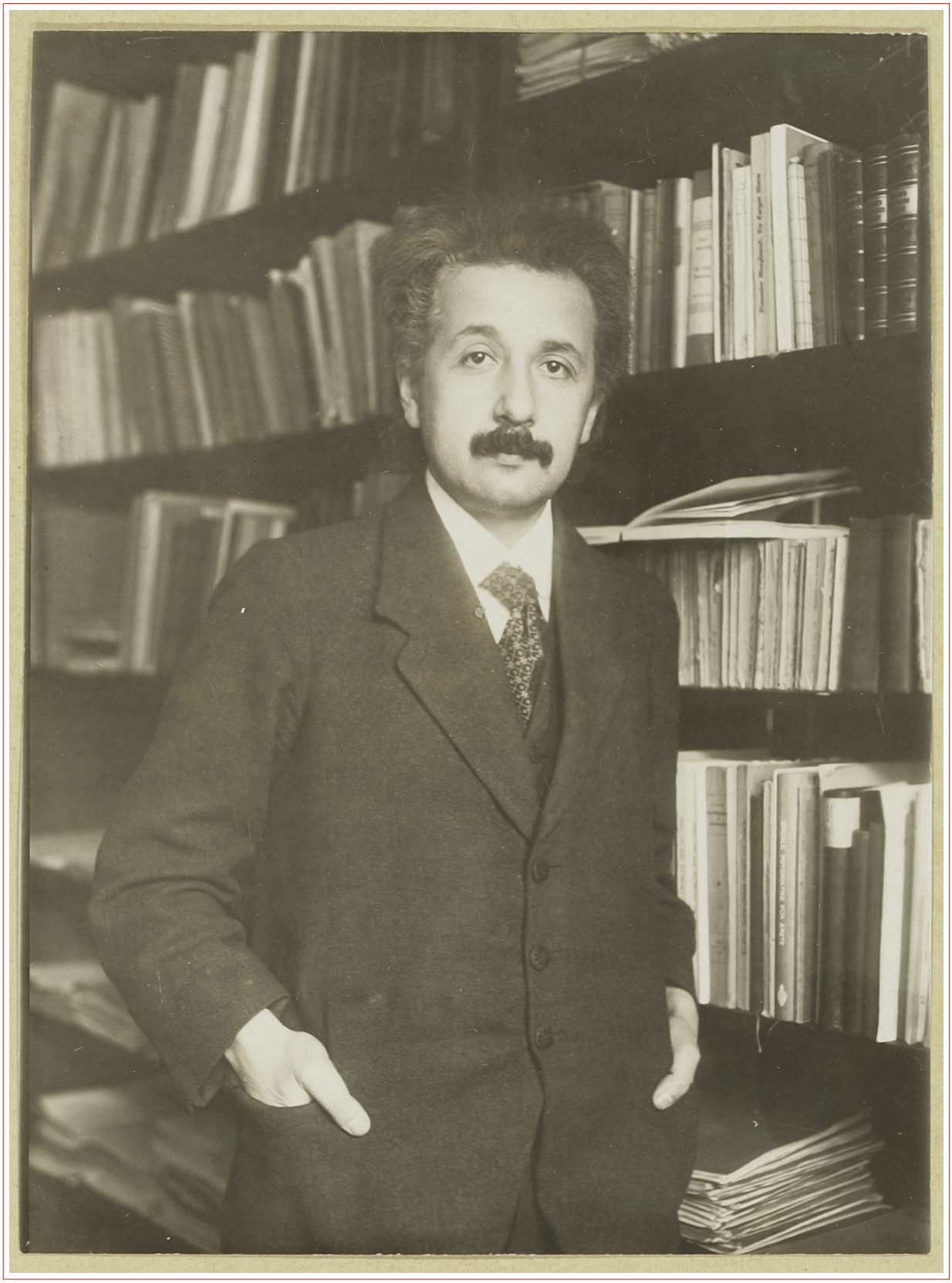
[67] 'Dieser Tensor G_{ik} ist der einzige Tensor, der für die Aufstellung allgemein kovarianter Gravitationsgleichungen zur Verfügung steht. Setzen wir nun fest, daß die Feldgleichungen der Gravitation lauten sollen $G_{\mu\nu} = -\kappa T_{\mu\nu}$, so haben wir damit allgemein kovariante Feldgleichungen gewonnen.' (Einstein, 1915b, p. 800). The Greek indices are in fact Einstein's own notation; he freely mixed these up with Latin ones.

[68] Footnote 1 in the November 18th paper (Einstein, 1915c) shows that Einstein knew the solution by then.

[69] See Einstein (1996a, Doc. 14, with extensive editorial notes on pp. 344–359.

[70] 'Denk Dir meine Freude beim Resultat, daß die Gleichungen die Perihel-Bewegungen Merkurs richtig liefern (...), Ich war einige Tage fassungslos vor Erregung.' From a letter to Ehrenfest, January 16, 1916, cf. Fölsing (1993, p. 418). Fokker (1955) also reports that Einstein had told him he had got palpitations after this computation.

[71] Like (1.20), the linearized form of (1.22) was already in the Zürich Notebook; the extra term $-\frac{1}{2}g_{\mu\nu}T$ also balances a corresponding term in the gravitational energy-momentum tensor (Janssen & Renn, 2020).



Albert Einstein in 1916 (Credit: Museum Boerhaave, Leiden)

1.7 Hilbert

Though this is not obvious from Einstein's papers, his mathematical colleague David Hilbert (1862–1943) played a significant role in the development of GR.[72] *Contra* his shallow and completely undeserved reputation of being a "formalist",[73] Hilbert was actually interested in physics throughout his career and often lectured on it. He combined this interest with his relentless emphasis on *axiomatization*, which started with his famous memoir *Grundlagen der Geometry* from 1899, in which he rewrote Euclidean geometry and heralded the modern era in mathematics. Exemplifying this, his *Sixth Problem* (from the famous list of 23 in 1900) reads:[74]

> Mathematical Treatment of the Axioms of Physics. The investigations on the foundations of geometry suggest the problem: To treat in the same manner, by means of axioms, those physical sciences in which already today mathematics plays an important part; in the first rank are the theory of probabilities and mechanics. (Hilbert, 1902a).

Hilbert's involvement with Einstein and relativity goes back to his joint seminar during the Winter Semester of 1907 at Göttingen with his friend and colleague Hermann Minkowski (1864–1909), who incidentally had been one of Einstein's teachers at Zürich (and did not think much of him). This seminar led Minkowski to his four-dimensional space-time view of special relativity, which after some hesitation also Einstein adopted and which of course was one of the keys to GR.

Hilbert was also interested in Mie's theory electromagnetic of matter from 1912, which has already been mentioned in connection with Einstein (1915b), and which was perhaps the first example of a "unified field theory". It was, in particular, based on an action principle (i.e. a Lagrangian, called a "world function" at the time), an idea which fitted well with Hilbert's notion of axiomatization and would play a central role in his work on gravitation to come.[75]

In 1915 Einstein came to Göttingen to give the Wolfskehl Lectures (from June 19 to July 7), which were devoted to general relativity and especially his *Entwurf Theorie*, which he still believed in at the time. Hilbert not only attended these lectures (as did e.g. Emmy Noether and Felix Klein), but he and his wife also hosted Einstein as their personal guest at home.[76] It seems that Einstein's visit triggered an all-out assault on the foundations of physics by Hilbert, who tried to combine elements of Mie's theory with Minkowski's space-time view of special relativit and Einstein's insights into the applicability of Riemannian geometry and the absolute differential calculus–all of which Hilbert was very familiar with–to the theory of gravitation.

[72] The sources for this subsection are Sauer (1999), Corry (2004) and Renn (2007), Vol. 4. The only biography of Hilbert is Reid (1970); a scientific biography is lacking. Rowe (2018) is a portrait of Hilbert's circle in Göttingen.

[73] Although he did not invent it, Hilbert was a pioneer of the view that rigorous mathematical proofs should be purely syntactic and hence independent of the meaning of the symbols in them, as long as the rules for manipulating these have been stated. This came to a head in the last part of his career (1920–1930), which was devoted to *Proof Theory* (a field of mathematics he did invent). But this kind of formalization was restricted to the analysis of proofs and axiom systems; until the 1920s Hilbert even stated axioms informally, combining mathematical and natural language. Outside this specific context, mathematics was as much alive for Hilbert as it is for anybody. A decade after he had played his role in GR, Hilbert also initiated the (serious) mathematical study of quantum mechanics, culminating in von Neumann's formalism based on Hilbert spaces; see e.g. Landsman (2022) and references therein.

[74] See e.g. Wightman (1976), Gorban (2018), and Corry (2018) for essays on Hilbert's Sixth Problem.

[75] Hilbert's interest in variational principles went back to his work on the Dirichlet principle (Hilbert, 1904).

[76] Corry (2004, p. 325) notes that Einstein and Hilbert had similar unconventional political views, notably their belief in the fundamentally international spirit of science. Neither had signed the patriotic and vitriolic manifesto *Aufruf an die Kulturwelt* from October 1914, in which 93 leading German intellectuals wholeheartedly supported the German side in the First World War. Among the signatories we find physicists like Fritz Haber and Max Planck, and mathematicians like Felix Klein; it was even more courageous of Hilbert not to sign it than it was of Einstein, since the former was a German citizen whereas the latter was, at the time, Swiss (though originally German by birth).

This led to two papers: Hilbert (1915, 1917), of which especially the first is of historical interest. Precisely because of the mix of the now outdated and partly incomprehensible Mie theory with Einstein's ideas (many of which have survived), Hilbert's reasoning is hard to follow. Even Einstein (who was familiar with Mie's theory and presumably also with his own) wrote:

> Why do you make it so hard for poor mortals by withholding the technique behind your ideas? It surely does not suffice for the thoughtful reader if, although able to verify the correctness of your equations, he cannot get a clear view of the overall plan of the analysis.[77]
> (Einstein to Hilbert, 30 May 1916)

Furthermore, there are serious discrepancies between the first galley proofs of Hilbert (1915) and the published version. These proofs presumably contain the paper in the form Hilbert presented it on 20 November 1915 in a formal colloquium to the Göttingen Academy of Sciences, i.e. five days before Einstein submitted his final paper (1915d) containing the correct field equations (this had even been preceded by an informal talk by Hilbert on 16 November with undoubtedly a similar content). On top of this, in a bizarre twist of events these galley proofs suffer from a deletion of precisely the part that *qua* location might have contained the correct field equations, cut out by an unknown person at an unknown time. Since Einstein and Hilbert were in regular contact during November 1915, this has led to wild speculations to the effect that Einstein had taken (or even *stolen*) his equations from Hilbert, or even if he hadn't, that Hilbert had at least scooped him and should get the credit for the invention of GR (seen as the "Einstein" equations). These speculations even stretched to the extent that Einstein fans had allegedly cut out the missing part because they *contained* the "Hilbert" equations–though by the same token Hilbert fans could have taken them out to hide the fact that they did *not contain* the Einstein equations.[78]

A detailed reconstruction shows that the parts missing from the galley proofs probably did *not* contain the correct field equations, or indeed any field equations, though they may well have contained the explicit Lagrangian $\sqrt{-g}R$, which may therefore correctly be called the *Hilbert Lagrangian*. Since Einstein (1916b) independently found this Lagrangian, but published it later than Hilbert (1915) even as published in 1916, the name *Einstein–Hilbert Lagrangian* is also correct. Having said this, knowing the correct Lagrangian, Hilbert could easily have derived the Einstein equations, for both the galley proofs and the published version of Hilbert (1915) show that he was perfectly familiar with the necessary variational techniques, and indeed all steps in the computation are indicated, except that the Lagrangian is left unspecified.

It seems that until 1916 Einstein was hardly influenced by the work of Hilbert (though he clearly admired him), except perhaps: (i) for his brief flirt with the electromagnetic world view in Einstein (1915b), which he discarded as soon as he could, and (ii) by Hilbert's competition speeding up his work in November 1915–since Hilbert sent him occasional updates on his work and invited him to at least his first talk in Göttingen on November 16, Einstein must have felt Hilbert breathing down his neck. On the other hand, the opposite influence is very clear from e.g. the differences between the galley proofs and the actual publication of Hilbert (1915).

[77] 'Warum machen Sie es dem armen Sterblichen so schwer, indem Sie ihm die Technik Ihres Denkens vorenthalten? Es genügt doch dem denkenden Leser nicht, wenn er zwar die Richtigkeit Ihrer Gleichungen verifizieren aber den Plan der ganzen Untersuchung nicht überschauen kann.' Quoted with translation by Stachel & Renn (2007), p. 881. This paper gives a detailed reconstruction and interpretation of Hilbert's work on GR.

[78] Before these galley proofs were discovered (by Corry in 2004), it was often suggested on fair grounds that Hilbert had priority over Einstein, since the published version of Hilbert (1915), which does contain the correct equations, carries a submission date of 20 November 1915. The tables initially turned when it was found that the galley proofs did not contain the Einstein equations, upon which the deleted section again complicated the issue. See Sauer (1999, 2005) and Rowe (2006) for a settlement and a review of the issue, respectively: which we follow.

These hectic events during November 1915 led to some tension between Einstein and Hilbert:

> The theory is beautiful beyond comparison. However, only *one* colleague has really understood it [i.e. Hilbert], and he is seeking to "*partake*" it (Abraham's expression) in a clever way. In my personal experience I have hardly come to know the wretchedness of mankind better than as a result of this theory and everything connected to it.[79] (Einstein to Zangger, 26 November 1915)

Hilbert did his best to alleviate the situation, for example by changing 'my theory' to 'the theory' in his galley proofs for Hilbert (1915), and by adding that the ten gravitational potentials $g_{\mu\nu}$ were 'first introduced by Einstein'. This helped, for a month later Einstein directly wrote him:

> There has been a certain ill-feeling between us, the cause of which I do not want to analyze. I have struggled against the feelings of bitterness attached to it, and this with complete success. I think of you again with unmarred friendliness and ask you to try to do the same with me. Objectively it is a shame when two real fellows who have extricated themselves somewhat from this shabby world do not afford each other mutual pleasure.
> With best regards, A. Einstein.[80] (Einstein to Hilbert, 20 December 1915)

What remains of lasting value, apart from his identification of the correct Lagrangian for GR, is Hilbert's recognition that the energy-momentum tensor $T_{\mu\nu}$ (which Einstein had to specify as such, even in cases where he relied on an action principle for the gravitational sector) is simply the variational derivative of the matter Lagrangian; although Hilbert only discovered this for electromagnetism, once the point had been made its generalization to other forms of matter was obvious. Of course, this made possible the derivation of the complete gravitational equations (including matter) from a single action principle, which Hilbert had been after all along.[81]

Furthermore, Hilbert was the first to use the (contracted) Bianchi identities in GR, deriving them (as we shall also do) from the invariance of the Ricci scalar under coordinate transformations (or diffeomorphisms), and drew the (highly nontrivial) conclusion that in electrodynamics the vacuum Maxwell equations $\nabla_\mu F^{\mu\nu} = 0$ follow from the coupling to gravity plus these Bianchi identities. See also §1.9. Finally, as many quotes (like the one opening this Introduction and the one ending §1.3) show, Hilbert quickly became a champion of GR, including Einstein's authorship of it (sometimes even at the expense of mentioning his own contributions). Coming from the leading mathematician in the world at a time in which Einstein was by no means yet the stellar figure he would later become, this undoubtedly helped the theory (and its creator).

[79] 'Die Theorie ist von unvergleichlicher Schönheit. Aber nur *ein* Kollege hat sie wirklich verstanden und sucht sie auf geschickte Weise zu "*nostrifizieren*" (Abraham'scher Ausdruck). Ich habe in meinen persönlichen Erfahrungen kaum je die Jämmerlichkeit der Menschen besser kennen gelernt wie gelegentlich dieser Theorie und was damit zusammenhängt.' Quoted with translation by Stachel & Renn (2007), p. 911. Heinrich Zangger (1874–1957) had been a friend of Einstein's since 1906. See Corry (2004, §9.2) on the culture of "nostrification" in Hilbert's Göttingen: 'It was widely understood, among German mathematicians at least, that "nostrification" encapsulated the peculiar style of creating and developing scientific ideas in Göttingen, and not least because of the pervasive influence of Hilbert. Of course, "nostrification" should not be understood as mere plagiarism.' (p. 419).

[80] 'Es ist zwischen uns eine gewisse Verstimmung gewesen, deren Ursache ich nicht analyseren will. Gegen das damit verbundene Gefühl der Bitterkeit habe ich gekämpft, und zwar mit vollständigem Erfolge. Ich gedenke Ihrer wieder in ungetrübter Freundlichkeit, und bitte Sie, dasselbe bei mir zu versuchen. Es ist objektiv schade, wenn zwei wirkliche Kerle, die sich aus dieser schäbiger Welt etwas herausgearbeitet haben, nicht gegensteitig zur Freude erreichen. Es grüsst Sie bestens, Ihr A. Einstein' (again taken from Stachel & Renn 2007, p. 913).

[81] In this context Lorentz (1916) should be mentioned, in which Hendrik Antoon Lorentz (1853–1928) develops a coordinate-free version of GR, based on a geometric interpretation of the Ricci scalar in the Lagrangian (Kox, 1988; Janssen, 1992). Unfortunately, since he did so just before the absolute differential calculus was geometrized by Levi-Civita's (1917a) invention of parallel transport (Iurato, 2016), his work on GR had very little influence.

1.8 Weyl

The bridge between Einstein's work and the modern mathematical approach to GR found in this book (and many others, including the great texts by Hawking & Ellis and by Misner, Thorne, & Wheeler, both of which appeared in 1973) is not so much Hilbert, whose mathematical style in GR is surprisingly old-fashioned and in fact hardly different from Einstein's, but his former PhD student Hermann Weyl (1885–1955). Weyl was an extraordinary broad and versatile mathematician, almost comparable with Hilbert himself, whose interest in physics as well as in the foundations of mathematics he also shared.[82] Weyl spent the years 1904–1913 in Göttingen,[83] where, clearly under the spell of Hilbert,[84] his early work was in functional analysis, then an upcoming field which was dominated at least in Germany by Hilbert's work on integral equations. Weyl's PhD thesis from 1908 was on singular integral equations and Fourier theory, after which his Habilitation thesis from 1910 was on Sturm–Liouville problems, seen in the context of what we now (since the work of von Neumann) call unbounded operators on Hilbert space.[85]

Weyl ended his Göttingen period with his famous book *Die Idee der Riemannsche Fläche* (1913), which launched the global study of Riemann surfaces and is one of the stepping stones towards to the modern definition of a manifold (see footnote 40). He then moved to Zürich (as the successor of Geiser, the man who had introduced Einstein to differential geometry), where he met Einstein and evidently got interested in relativity. In 1918 Weyl published his lecture notes *Raum - Zeit - Materie (Space - Time - Matter)*, from which we already quoted the preface at the beginning of this historical overview. Einstein himself wrote a glowing review:

> I am always tempted to read the individual parts of this book again, because every page shows the amazingly steady hand of the master who has penetrated the subject matter from the most diverse angles. I consider it a happy occasion that such a distinguished mathematician has taken care of this new field. He understood how to combine mathematical rigor with graphic intuition. From this book, the physicist can learn the foundations of geometry and the theory of invariants, and the mathematician can learn those of electricity and the theory of gravitation. (...) One especially sees there with amazement how the most complicated becomes simple and self-evident under *Weyl's* hand. (...) It is here that *Weyl* not only demonstrates his easy mastery of the mathematical form, but also his deep insight into what is essential in physics. (...) The expositions of the last paragraphs exemplify how a *born* mathematician can be effective here through simplifying and clarifying. The book will be invaluably helpful to everybody who wants to work in this field, not to mention the pure joy derived from its study.[86] (Einstein, 2002, pp. 62–63)

[82] The latter interest even led to a break between them at the time when Weyl supported Brouwer's intuitionism.

[83] See e.g. Eckes (2019). There seems to be no biography of Weyl, but see Scholz (2001) for his mathematics and especially *Raum - Zeit - Materie*, and Ryckman (2005) for his philosophy (mostly in connection with GR).

[84] 'One cannot overstate the significance of the influence exerted by Hilbert's thought and personality on all who came out of [the Mathematical Institute at Göttingen]' (Corry, 2018). However, Eckes (2019) draws attention to the considerable influence that also Zermelo and Klein (and perhaps also Minkowski) had on the young Weyl.

[85] The *limit point - limit circle theorem* from his Habilitation thesis is still used.

[86] 'Immer wieder drängt es mich dazu, die einzelnen dieses Buches von neuem durchzulesen: denn jede Seite zeigt die unerhört sichere Hand des Meisters, der den Gegenstand von den verschiedensten Seiten durchdrungen hat. Ich betrachte es als einen Glücklichen Umstand, daß ein so ausgezeichneter Mathematiker sich des neuen Gebiets angenommen hat. Er hat es verstanden, mathematische Strenge mit Anschaulichkeit zu verbinden. Der Physiker kann aus seinem Buche die Grundlagen der Geometrie und Invarianzentheorie, der Mathematiker diejenigen der Elektrizitätslehre and Gravitationstheorie lernen. (...) Hier sieht man ganz besonders mit Staunen, wie in *Weyls* Händen das Kompliziertesteeinfach und selbstverständlich wird. (...) Hier zeigt sich so recht, daß *Weyl* nicht

Weyl's book is remarkable in many ways, including its attractive mix of mathematics, physics, and philosophy,[87] but also its lyrical–if not, occasionally, outright hysterical–prose, which is very unusual for a mathematical physics text. For example, the fourth edition ends as follows:[88]

> Whoever looks back over the ground that has been traversed (...) must be overwhelmed by a feeling of freedom won–the mind has cast off the fetters which have held it captive. He must feel transfused with the conviction that reason is not only a human, a too human, makeshift in the struggle for existence, but that, in spite of all the disappointments and errors, it is yet able to follow the intelligence which has planned the world, and that the consciousness of each one of us is the centre at which the One Light and Life of Truth comprehends itself in Phenomena. Our ears have caught a few of the fundamental chords from that harmony of the spheres of which Pythagoras and Kepler once dreamed.[89] (Weyl, 1921, p. 284)

Back on earth, Weyl was the first author to describe tensors in a coordinate-free manner as multilinear maps, even starting the technical part of his book with an axiomatic treatment of vector spaces.[90] He then defines tensors as pointwise multilinear maps, just as we do; although in the spirit of the time Weyl uses coordinates as soon as he can, the abstract underpinning is clearly there. His most significant mathematical innovation was the idea of an *affine connection* (cf. our §3.3, where it is called a *linear* connection), which–though somewhat paradoxically introduced through old-fashioned infinitesimals–gives a covariant derivative (as well as the associate notion of parallel transport) *independently of the metric*.[91] Assuming the affine connection to be torsion-free (for which he gives some arguments), Weyl also proves that if there is a (nondegenerate) metric, what we now call the Levi-Civita or metric connection is the unique affine connection for which parallel transport preserves length. His derivation of the Einstein equations from an action principle follows Hilbert, including the definition of the energy-momentum tensor as the variational derivative of the matter action with respect to the metric.

However, arguably the most lasting contribution of *Raum - Zeit - Materie* (from the third edition onwards) is Weyl's idea of (conformal) *gauge symmetry*, taken up in the next section.

---

nur die mathematische Form spielend meistert, sondern auch mit tiefem Blick für das physikalische Wesentliche begabt ist. (...) Die Darlegungen der letzten Paragraphen zeigen, wie vereinfachend und klärend der *geborene* Mathematiker da wirken kann. Jedem, der an dem Gebiet mitarbeiten will, wird das Buch unschätzbare Dienste leisten, abgesehen von der reinen Freude, die er beim Studium findet.' (Einstein, 1918b).

[87] Weyl's wife, Helene (1893–1948), whom he incidentally betrayed with Schrödinger's wife when they were all in Zürich from 1921–1927 (and Weyl helped Schrödinger with the solution of the equation named after him), was a student of Edmund Husserl and an intellectual in her own right. Weyl himself also tended towards phenomenology.

[88] There are many editions of the book, of which the first (1918) and the second (1919) are identical. The third (1919) and the fourth (1921) editions are major updates, especially the third, in which Weyl introduces his own idea of an affine connection without having a metric. The English translation from 1922 is from the fourth edition.

[89] 'Wer auf den durchmessenen Weg zurückschaut (...) muß von dem Gefühl errungener Freiheit überwältigt werden–ein festgefügter Käfig, in den das Denken bisher gebannt war, ist gesprengt–; er muß durchdrungen werden von der gewißheit, daß unsere Vernunft nicht bloß ein menschlicher, allzumenschlicher Notbehelf im Kampf des Daseins, sondern ungeachtet alle Trübungen and alles Irrtums doch der Weltvernunft gewachsen ist und das Bewußtseins eines jeden von uns der Ort, wo das Eine Licht und Leben der Wahrheit sich selbst in der Erscheinung ergreift. Ein paar Grundakkorde jener Harmonie der Sphären sind in unser Ohr gefallen, von der Pythagoras und Kepler träumten.' Translation: Henry L. Brose, pp. 311–312 in Weyl (1922).

[90] The lack of references in §I.2 is an example of the Göttingen habit of "nostrification" (cf. footnote 79), since the axioms had already been given by Peano in 1888 (Moore, 1995). It is unclear whether Weyl knew Peano's work.

[91] Compared with Levi-Civita (1917a), this makes an ambient flat space unnecessary even in the metric case.

1.9 Mathematical foundations of GR: Towards the modern era

The full history of post-1915 GR remains to be written, even on the physics side. Einstein himself continued to make major contributions to his theory, among which perhaps Einstein (1917b), the paper that launched relativistic cosmology (and introduced his cosmological constant), and Einstein (1918c), in which he predicted gravitational waves (directly detected almost a century later, on 14 September 2015), stand out.[92] Moreover, the confirmation of the general relativistic prediction of the gravitational bending of sunlight, announced by Eddington in a session of the Royal Society on November 6, 1919, sanctified by J.J. Thomson in the Chair as 'the most important result obtained in connection with the theory of gravitation since Newton's day', and picked up by the world press, made Einstein the celebrity that he has remained until the present.

Nonetheless, despite the undeniable power and beauty of the theory and the increasing fame and prestige of its creator, GR remained at least in physics a niche field until the 1960s. It was immediately picked up by the leading astronomer of the day, Eddington, as well as by his (now almost equally famous) colleagues De Sitter, and Lemaître, similarly by the greatest mathematician of his era, Hilbert, as already mentioned, followed by Levi-Civita, Weyl, and (Élie) Cartan in his footsteps. Even major philosophers like Cassirer, Reichenbach, and Schlick wrote about the implications of GR.[93] However, with a few exceptions the response from the physics community (that Einstein himself–never a real astronomer, mathematician, or philosopher–came from!) was lukewarm at best.[94] This attitude may have been partly due to the hostility to German science during and after World War I (although Einstein, while residing in Berlin, had renounced his German citizenship as early as 1896 and was a Swiss citizen at the time). Not coincidentally, Eddington and, from the other side, Hilbert were among the very few academics who were interested in overcoming this hostility. But it lasted for decades. On a different note, Ehlers (2007) writes: 'At that time [the late 1940s] general relativity was considered a difficult and useless subject, admitting no interaction between theory and experiment.' Or (Bryce) DeWitt:

> Most of you can have no idea how hostile the physics community was, in those days, to persons who studied general relativity. It was worse than the hostility emanating from some quarters today towards the string-theory community. In the mid fifties, Sam Goudsmit, then Editor-in-Chief of the *Physical Review* and *Physical Review Letters*, would no longer accept "papers on gravitation or other fundamental theory." (DeWitt-Morette, 2011, p. 6)

The first international conference on general relativity was only held in 1955, and its subsequent revival was due to a small group of dedicated people, partly inspired by applications to astrophysics and cosmology, and partly for the theory's own sake.[95] This led to important GR communities in the United States (Bergmann, DeWitt, Schild, especially Wheeler at Princeton), the Soviet Union (Fock, Ivanenko, especially Zeldovich in Moscow), and Europe, e.g. in France (Lichnerowicz, Choquet-Bruhat), Germany (Jordan), Poland (Infeld, Trautman), Ireland (Lanczos, Synge, Schrödinger), and the United Kingdom (Bondi, Dirac, Hoyle, McCrea, Whitrow, Penrose).[96] In particular, Dirac's student Sciama created the GR school at Cambridge that still exists today and once included Hawking, Carter, Ellis, Rees, and many other leading relativists.

[92] See Jannsen & Lehner (2014), and of course *The Collected Papers of Albert Einstein*, from Volume 6 onwards.

[93] See Ryckman (2005) for the reception of general relativity among philosophers.

[94] An exception is Pauli (1921), which he wrote at the age of 21 at the behest of his mentor Sommerfeld. This was the first complete review of GR after Weyl. Another exception was Einstein's friend Lorentz, see footnote 81.

[95] See Thorne (1994), Kaiser (1998), Eisenstaedt (2006), Melia (2009), Ashtekar (2014), Blum, Lalli, & Renn (2015, 2016, 2020), Goenner (2017), and Lalli (2017) for personal and scholarly historical studies of this.

[96] See Robinson (2019) about King's College London, and Lalli (2017) for smaller GR groups since the 1950s.

In view of the overall structure of this book, in the remainder of this section we restrict ourselves to a few brief comments about the transition from what was known to say Hilbert and Weyl around 1917, to the current formalism of mathematical GR. It would be fair to say that Hilbert mainly looked at GR from the point of view of PDEs, whereas Weyl had a more geometric view, which he combined with an emphasis on causal structure, as explained below. These different perspectives initially developed separately, in that the causal theory did not rely on the PDE theory whilst the initial PDE results were local in nature. But the two areas meet through the absolutely central notion of global hyperbolicity that is common to both, and in modern mathematical GR they are inseparable (although one still has specialists on either side). We first discus the PDE approach (which may indeed be a few months older than the causal one).

Hilbert (1917) predated Hadamard (1923), in which the Cauchy problem for PDEs was first stated. The solution to a given PDE should: (i) *exist* on a given domain for all suitable initial and/or boundary data;[97] (ii) be *uniquely determined* by these data, and (iii) be *stable* against variations in these data (typically as expressed by continuity with respect to certain norm-topologies). This was seen as the form of determinism (or "causality") appropriate for physics. In 1917 the second volume of Courant & Hilbert (1937), which gave a complete treatment of PDEs as the field was known at the time, was also twenty years in the future.[98]

However, in 1917 Hilbert certainly possessed massive knowledge of nineteenth century PDE theory, as well as of the early twentieth century interaction between PDEs and functional analysis, of which field he had been one of the founders. In particular, Hilbert recognized that Einstein's equations were not of any standard type (i.e. hyperbolic, elliptic, or parabolic) and that because of what we now call "Bianchi identities" their initial value problem was ill-posed in the sense that reasonable initial data do not determine a unique solution; cf. §1.5. He foresaw what we now call *geometric uniqueness*, see Theorem 7.8 in §7.6, in stating that 'physically meaningful' quantities *were* uniquely determined, and that using suitable coordinates (namely geodesic normal coordinates, which he called 'Gaussian') also led to uniqueness in general.[99]

The next important contribution to the PDE side of GR was made by Emmy Noether (1882–1935), whose famous article 'Invariante Variationsprobleme' explained the difficulties with Einstein's equations that Hilbert had found in terms of the infinite-dimensional symmetries of the action or Lagrangian from which these equations could be derived, and introduced what are now called the first and second Noether theorems.[100] However, her paper is so general that, despite a final section commenting on Hilbert's work, it does not contain any detailed expressions for GR.

In that sense, it was Georges Darmois (1888–1960), who, citing neither Hilbert nor Noether, (co) founded the theory of the constraints of GR. Darmois (1927) recognized the equations

$$G_{\mu 0} = 0 \tag{1.23}$$

[97] For hyperbolic PDEs such as the wave equation one has initial data; for elliptic PDEs like Laplace's equatio one has boundary data; and for parabolic PDEs such as the heat equation one has combinations thereof.

[98] It is impossible to resist quoting a piece from Weyl's review of this is book, which though entirely written by Courant clearly carried Hilbert's spirit: 'Nowadays many mathematical books do not seem to be written by living men who not only know, but doubt and ask and guess, who see details in their true perspective–light surrounded by darkness–who, endowed with a limited memory, in the twilight of questioning, discovery, and resignation, weave a connected pattern, imperfect but growing, and colored by infinite gradations of significance. The books of the type I refer to are rather like slot machines which fire at you for the price you pay a medley of axioms, definitions, lemmas, and theorems, and then remain numb and dead however you shake them.' (Weyl, 1938, p. 602).

[99] See Stachel (1992) for a more detailed analysis of Hilbert's contribution, as well as for the history of the Cauchy problem of GR up to the work of Choquet-Bruhat. For general PDE history, see Brezis & Browder (1998).

[100] The original source is Noether (1918). A sample of the extensive secondary literature is Kossmann-Schwarzbach (2011, 2020), Eggertsson (2019), and Read, Teh, & Roberts (2021). Rowe (2021) is a biography of Noether.

as conditions restricting the initial data. He saw that they automatically "propagate" (in holding everywhere provided they are satisfied at $t = 0$ and the other equations hold, see §7.5), and also gave their geometric expressions (7.148) - (7.149) in terms of the first and second fundamental forms of the embedded Cauchy surface where they are imposed. He also showed that in the wave gauge (or harmonic gauge, first used by De Donder) the remaining six Einstein equations were hyperbolic propagation equations for each of the components $g_{\mu\nu}$ of the metric. Finally, he studied the possibility of giving initial data on null surfaces (i.e. lightcones), in which he was far ahead of his time. This is very impressive for someone who was actually a statistician!

Darmois was also the thesis advisor of André Lichnerowicz (1915–1998), who worked in GR from 1937 until 1967 and made many contributions to the field.[101] His most important work, collected in Lichnerowicz (1955), includes his theorem on asymptotically flat space-times (see §8.4), as well as his conformal analysis of the constraint equations (see §8.6). His importance as an organizer of the French GR community, e.g. through organizing the *Journées Relativistes* conference series, can hardly be overestimated.[102] In that capacity Lichnerowicz was also the PhD advisor of Yvonne Choquet-Bruhat (born in 1923), who, during a career that spanned sixty years, from a four-page announcement (Fourès-Bruhat, 1948) to a comprehensive 800-page textbook *General Relativity and the Einstein Equations* (Choquet-Bruhat, 2009), led the PDE approach to GR by giving direction and proving two of the most important results herself, namely the first local existence and uniqueness result (Fourès-Bruhat, 1952) and the crowning maximal existence and uniqueness theorem, which she proved in 1969 with Geroch.[103]

Subsequent work on the PDE aspects of GR falls into two directions, which might be called *hyperbolic* and *elliptic*, depending on whether one works mainly on the evolution equations or on the constraint equations, respectively, or, phrased differently, on the evolution of the initial data or on the initial data themselves. Of course, these aspects cannot be entirely separated.

On the hyperbolic side, one studies global properties of the above maximal (globally hyperbolic) solutions, notably their *extendibility* (which does not contradict the formal property of maximality) and *stability*. Even the simplest case, namely the question of the stability of Minkowski space-time under small perturbations of its initial data, took a 500+ page *book* (by Christodoulou and Klainerman) to settle it in the positive. Despite later simplifications of this proof, analogous current work on the stability of black holes solutions is published in *papers* whose page count even runs over 800. Apart from stability problems, other goals of this approach include (dis)proving Penrose's cosmic censorship and final state conjectures (see chapter 10).

On the elliptic side, one highlight has been the proof of the *positive mass theorem* by Schoen & Yau (1979) and Witten (1981), to which a brief introduction will be given in §8.4.[104] Many of the techniques used in proving uniqueness or "no hair" theorems for black holes (see §§10.9 - 10.10) also come from the elliptic approach. Another achievement has been the development of *gluing techniques* for solutions to the Einstein equations by gluing their initial data.[105]

[101] See Lichnerowicz (1992) for a brief memoir, in which he pays special tribute to Élie Cartan (1869–1951), one of the founders of modern Lie theory and differential geometry, who also did important work motivated by GR, including the geometric reformulation of Newtonian gravity now called Newton–Cartan theory (Malament, 2012).

[102] In this respect also the *Les Houches* schools founded in 1951 by Cécile Morette should be mentioned.

[103] The historical survey by Ringström (2015) explains the precise regularity of the solutions in Fourès-Bruhat (1952) and also puts her work in a much wider mathematical perspective. Choquet-Bruhat (2014) also looks back on her results; see also her autobiography *A Lady Mathematician in this Strange universe* (Choquet-Bruhat, 2018).

[104] Roughly speaking, in any asymptotically flat space-time (i.e. one in which the metric approaches the Minkowski metric at infinity) one can define a quantity in terms of the metric, which for the Schwarzschild solution is the mass of the star (or black hole), but which in general is not obviously positive. The theorem states that it *is* positive.

[105] See e.g. Chruściel, Galloway, & Pollack (2010), which is actually a general survey of mathematical GR.

We now turn to the "causal" approach to GR. Its characteristic emphasis on the conformal structure of GR, i.e. the equivalence class of the metric tensor g under a rescaling

$$g_{\mu\nu}(x) \mapsto e^{\lambda(x)} g_{\mu\nu}(x), \tag{1.24}$$

with λ an arbitrary smooth function of space and time, originated with Weyl (1918b). Although he mentions the analogy with Riemann surfaces,[106] which undoubtedly drove him in this direction, his real argument is that what he calls *Reine Infinitesimalgeometrie* must go beyond Riemannian geometry, which (according to Weyl) suffers from the defect that parallel transport of vectors (through the metric or Levi-Civita connection, a concept Weyl himself had co-invented) preserves their length. This makes length of vectors an absolute quantity, which a 'pure infinitesimal geometry' or a theory of *general* relativity cannot tolerate. To remedy this, Weyl introduced the idea of *gauge invariance*, stating that the laws of nature should be invariant under the rescaling (1.24). To this end, he introduced what we now call a gauge field $\varphi = \varphi_\mu dx^\mu$ and a compensating transformation $\varphi_\mu(x) \mapsto \varphi_\mu(x) - \partial_\mu \lambda(x)$, and identified φ with the electromagnetic potential (i.e. A). Dancing to the music of time, he then proposed that the pair (g, φ) describes all of physics. This is not the case,[107] but the idea of gauge symmetry has lasted and forms one of the keys to modern high-energy physics and quantum field theory: serpenditiously, although it is misplaced in the *classical gravitational* context in which Weyl proposed it, through the Standard Model it has ironically become a cornerstone of *non-gravitational quantum* physics!

The conformal structure of a Lorentzian manifold determines the lightcones (and hence also their interiors), and as such Weyl was of course not the only author to discuss causal structure. For example, Einstein (1918c) himself wondered if gravitational wave propagate with the speed of light, and showed this in a linear approximation; Weyl mentions this also.[108] Furthermore, independently of Weyl, and in fact inspired by special rather than general relativity, Robb (1914, 1936), Reichenbach (1924), Zeeman (1964), and also others axiomatized causal structure as a specific *partial order*. In modern notation, if M is Minkowski space-time then the simplest such relation is $J^+ \subset M \times M$, where $(x,y) \in J^+$ or $x \leqslant y$ if y lies within or on the future lightcone emanating from x. For general relativistic space-times this may be generalized by defining $(x,y) \in J^+$ iff there exists a future-directed causal curve from x to y (see §5.3).

These themes–gravitational radiation, conformal invariance, and causal order, with additional inspiration from some of the drawings of the Dutch artist M.C. Escher–were combined and came to a head in the work of Roger Penrose (born in 1931). Between 1963 and 1972, with a last eruption in 1979 (see chapter 10), Penrose introduced the global causal techniques and ideas in GR that are now central to the mathematical analysis of the subject.[109] Moreover, in 1965 he used these techniques to prove the first singularity theorem of GR, based on his concept of a trapped surface.[110] This inspired the singularity theorem of Stephen Hawking (1942–2018), whose Adams Prize Essay (Hawking, 1966), along with the book by Hawking & Ellis (1973) that arose from it, may also be counted among the founding documents of mathematical GR.[111]

[106] See page 397. Riemann surfaces may equivalently be defined as either one-dimensional complex manifolds or as two-dimensional Riemannian manifolds *up to conformal equivalence* Modestly, Weyl does not cite his own decisive contribution to their theory (Weyl, 1913). This equivalence also influenced Penrose's work on GR.

[107] See Einstein's negative reaction to Weyl (1918b) in Einstein (2002a), Doc. 8. See also Goenner (2004), §4.1.3.

[108] See e.g. page 251 of the English translation of the fourth edition of *Raum - Zeit - Materie* (Weyl, 1922).

[109] A key exception is the notion of global hyperbolicity, which has its roots in the work of Leray (1953) and was adapted to GR by Choquet-Bruhat (1967) and Geroch (1970). Penrose (1963) also worked on the PDE side.

[110] See footnote 270 for references on the history of the singularity theorems. See also chapter 6.

[111] See Ellis (2014) for the historical context of this essay and of Hawking's early work in general.

1.10 Epilogue: General covariance and general relativity

In this appendix to our historical introduction we return to the theme of general covariance and its possible relationship to some relativity principle that generalizes the one underlying Einstein's special theory of relativity. Starting with Einstein himself, this issue has naturally concerned many people, without a clear conclusion. But one may at least try to avoid some pitfalls.[112]

Although the field equations in Einstein (1915d) were generally covariant at last, it took Einstein another year to relieve himself of all coordinate conditions. Einstein (1916a) still gives the vacuum field equations in the form $\tilde{R}_{\mu\nu} = 0$ under the unimodular coordinate condition (1.21), and also their derivation from an action principle is the same as the one he gave in the previous year (Einstein, 1915a). It is only in Einstein (1916b), where he derives the generally covariant equations (1.22) from what we now call the Einstein–Hilbert action, that we read:

> On the other hand, in antithesis to my own most recent treatment of the subject, there is to be complete liberty in the choice of the system of co-ordinates.[113]

This, then, is what Einstein meant by "general covariance". But he also believed that generally covariance implies that GR satisfies a "general principe of relativity". Returning to the reconstruction of his reasoning in §1.3, there is little doubt that in conflating symmetry properties of GR as a whole with symmetry properties of its solutions Einstein actually cornered himself:

- Either he explains why in GR geodesic frames of reference are equivalent to arbitrary frames. But then the same argument (whatever it is) would apply to Minkowski space-time, and he loses the perfect match between the special principle of relativity and the special theory of relativity, on which his arguments for general relativity were predicated.
- Or he accepts that geodesic frames of reference are preferred (i.e. "special") and hence blasts the general principle of relativity even in GR. *Every way you look at it you lose!*

In fact, the difference between the theories of special and general relativity cannot lie in general covariance. Consider, in Einstein's own language, the following two equations for the metric:

$$R_{\mu\nu} = 0; \tag{1.25}$$

$$R^{\rho}_{\sigma\mu\nu} = 0, \tag{1.26}$$

or, in modern notation, $\mathrm{Ric}(g) = 0$ and $\mathrm{Riem}(g) = 0$, respectively, where the former is the Ricci tensor, the latter is the Riemann tensor, and g is a Lorentzian metric to be solved for. Eq. (1.25) are the vacuum *Einstein equations*, and let us call (1.26) the *Minkowski equations*. They share exactly the same covariance properties, but (1.25) gives GR (without matter) whereas by a basic result in Riemannian geometry (1.26) gives special relativity.[114] More generally, almost any physical theory of the kind known in classical mechanics and field theory can be geometrized and, at the expense of adding equations like (1.26), be made generally covariant. Thus Einstein faces two problems in trying to relate general covariance to general relativity (i.e. of motion):

[112] The history of the debate on general covariance is reviewed in Norton (1993, 1995). Further literature includes Anderson (1967), Friedman (1983), Norton (1989, 1999), Brown (2005), Dieks (2006), Earman (2006ab), Giulini (2007), Pooley (2015), Wallace (2017), and Dewar (2020). Though closely related, we do not enter the philosophical debate between *substantivalism* and *relationalism*, which was revisited in the light of the hole argument and general covariance by Earman & Norton (1987) and Butterfield (1987, 1989). See also Pooley (2017, 2020) for reviews.

[113] 'Anderseits soll im Gegensatz zu meiner eigenen letzten Behandlung des Gegenstandes die Wahl des Koordinatensystems vollkommen freibleiben.' (Einstein, 1916b, p. 1111). Translation by W. Perrett and G.B. Jeffery.

[114] At least locally, see Theorem 4.1. Eq. (1.26) is equivalent to (1.25) plus the vanishing of the Weyl tensor.

1. GR distinguishes geodesic motion from any other and hence *does* contain preferred frames. We may then ask what kind of relativity (if not relativity of motion) GR *does* generalize.

2. Almost any physical theory can be made generally covariant. Thus general covariance cannot be equated with general relativity and by itself must be physically empty.[115] This raises the question which physical property, if any, general covariance *does* express.

As to the first point, we are more optimistic than the great relativist Synge, who wrote:

> The name [general theory of relativity] is repellent. Relativity? I have never been able to understand what that word means in this connection. I used to think that this was my fault, some flaw in my intelligence, but it is now apparent that nobody ever understood it, probably not even Einstein himself. So let it go. What is before us is Einstein's theory of gravitation.[116] (Synge, 1966, p. 7)

However, in developing his special theory Einstein used the notion of "relativity" in a way that does seem to survive into GR in a defensible way. In special relativity, apart from the broad–one would like to, but cannot say "general"!–relativity principle stating that the laws of physics are the same in each inertial frame, in the context of electrodynamics that actually led him to his theory, Einstein (1905) made the following point, with which he even starts:

> It is known that Maxwell's electrodynamics–as usually understood at the present time–when applied to moving bodies, leads to asymmetries which do not appear to be inherent in the phenomena. Take, for example, the reciprocal electrodynamic action of a magnet and a conductor. The observable phenomenon here depends only on the relative motion of the conductor and the magnet, whereas the customary view draws a sharp distinction between the two cases in which either the one or the other of these bodies is in motion. For if the magnet is in motion and the conductor at rest, there arises in the neighbourhood of the magnet an electric field with a certain definite energy, producing a current at the places where parts of the conductor are situated. But if the magnet is stationary and the conductor in motion, no electric field arises in the neighbourhood of the magnet. In the conductor, however, we find an electromotive force (...) which gives rise (...) to electric currents of the same path and intensity as those produced by the electric forces in the former case.[117] (Einstein, 1905, p. 891)

[115] This was first pointed out to Einstein in a (now) famous paper by a high-school teacher called Kretschmann (1917). Einstein grudgingly conceded this point; see Norton (1993) and Giovanelli (2013, 2019) for a study of their debate. Kretschmann raised his concerns in the specific context of Einstein's "point-coincidence argument", which was Einstein's answer to his earlier "hole argument" discussed in §1.5: which had led him to temporarily abandon general covariance, almost blocking his way to GR. Thus Kretschmann argued that any physical theory whose empirical content lies solely in point-coincidences (as Einstein had it) can be written in generally covariant form.

[116] This quotation has been borrowed from Norton (1995). John Lighton Synge (1897–1995) wrote powerfully and beautifully. The entire preface of his book (Synge, 1966) would be worth quoting, or at least the brilliant first page.

[117] 'Daß die Elektrodynamik Maxwells - wie dieselbe gegenwärtig aufgefaßt zu werden pflegt–in ihrer Anwendung auf bewegte Körper zu Asymmetrien führt, welche den Phänomenen nicht anzuhaften scheinen, ist bekannt. Man denke z. B. an die elektrodynamische Wechselwirkung zwischen einem Magneten und einem Leiter. Das beobachtbare Phänomen hängt hier nur ab von der Relativbewegung von Leiter und Magnet, während nach der üblichen Auffassung die beiden Fälle, daß der eine oder der andere dieser Körper der bewegte sei, streng voneinander zu trennen sind. Bewegt sich nämlich der Magnet und ruht der Leiter, so entsteht in der Umgebung des Magneten ein elektrisches Feld von gewissem Energiewerte, welches an den Orten, wo sich Teile des Leiters befinden, einen Strom erzeugt. Ruht aber der Magnet und bewegt sich der Leiter, so entsteht in der Umgebung des Magneten kein elektrisches Feld, dagegen im Leiter eine elektromotorische Kraft (...) die aber (...) zu elektrischen Stromen von derselben Größe und demselben Verlaufe Veranlassung gibt, wie im ersten Falle die elektrischen Kräfte.' Translation by W. Perrett and G.B. Jeffery (Einstein *et al.*, 1923). In connection with GR see also Janssen (2012, 2014).

In modern language, the separation of the electromagnetic field $F_{\mu\nu}$ into an electric part F_{0i} and a magnetic part F_{ij} depends on the observer. Similarly, if in GR one identifies the metric $g_{\mu\nu}$ with the frame-independent "inertio-gravitational potential" and the Christoffel symbols $\Gamma^{\rho}_{\mu\nu}$ with the actual gravitational field, then a freely falling person locally puts $\Gamma^{\rho}_{\mu\nu}$ to zero and hence feels no gravitational field, whereas a stationary observer has non-vanishing Christoffel symbols and hence attributes the observed motion to gravity. From this point of view, the non-tensorial character of the Christoffel symbols is a coordinate-dependent blessing in disguise!

Another concept that perhaps GR relativizes more generally than special relativity does is *simultaneity*. More than anything else, what makes special relativity a genuine challenge to our world view is that the *now* has become (inter)subjective: for an observer at rest in the usual $(t,\vec{x})$ coordinates the planes of simultaneity are horizontal, whereas for a (relatively) moving observer they are tilted–*although they are still planes* (this explains phenomena like length contraction and time dilation). As we shall see in chapter 8, in the $3+1$ split of GR there is no preference for a foliation of space-time by "horizontal" or "flat" planes; essentially any choice of hypersurfaces of simultaneity, typically curved, is allowed. See also §8.11.

The second question remains. Conceding that any physical theory could indeed be brought into generally covariant form using the absolute differential calculus, Einstein's own answer was that only some theories (including, of course, GR) are 'simple and transparent' in generally covariant form.[118] With due respect,[119] this is balderdash. First, eq. (1.26) is as simple and transparent as (1.25), or even simpler, since (1.25) is a contraction of (1.26). Also, the example of Newtonian gravity that Einstein gave would soon be reformulated in geometric fashion by Cartan, resulting in a theory as simple and transparent as GR. Second, criteria like simplicity, transparency, and beauty are subjective, time-dependent, and relative to one's (mathematical) education. In 1900 only a few mathematicians and physicists were familiar with linear algebra, but now this is a first-year subject which most students find simpler than, say, analysis.[120]

Another answer is that GR "is a geometric theory". But this is not a very good answer either, since special relativity is as geometric (and as covariant) as GR, as we have already seen. In fact, Einstein himself was not very impressed by this argument at all, pointing out that any theory containing vectors (which would mean practically all of physics) could be called "geometric".[121]

A third answer would be that general covariance expresses the fact that GR "lacks absolute objects".[122] But once again, formulated like (1.26), so does special relativity, at least in writing down its equations. Perhaps it ends up with an absolute object, namely the Minkowski metric, but then again, any specific solution to (1.25) is also an absolute object in the same sense.[123]

[118] See Einstein's (1918a) answer to Kretschmann (1917) as well as his 1954 letter to De Broglie (cf. §1.3).

[119] Einstein's (later) views in this respect, for which Norton (2000) presents a historical analysis, were similar to Dirac's, see e.g. the book chapter 'Mathematical Beauty' by his biographer H. Kragh, `https://simplydirac.pressbooks.com/chapter/mathematical-beauty/`. See also Hossenfelder (2018).

[120] This objection also applies to various refinements of Einstein's point that are discussed by Norton (1993, 1995). For example, Bergmann (1942) argued that GR stands out because, starting from a generally covariant reformulation, the structure of other theories (like special relativity and Newtonian gravity) simplifies if their covariance group is reduced. Here again one wonders what "simplicity" means: if it means "less structure", then special relativity would be simpler in its generally covariant form, whose equations assume just a metric, rather than a specific one.

[121] See Lehmkuhl (2014).

[122] The idea is that the symmetry group is the largest one preserving all "absolute objects". This works for special relativity if (only) the Minkowski metric is seen as absolute, and it works for GR if nothing is deemed absolute.

[123] The difference between GR and generally covariant special relativity is that the latter is *categorical*, in–barring the topology of space-time–having only one solution, up to isomorphism, much as Hilbert's (1900) axioms for the real numbers (as a complete totally ordered field) are categorical, at least if they are expressed in second-order logic (Shapiro, 1991). Whatever its implications, categoricity does not hold for e.g. Newton–Cartan gravity.

Moreover, Minkowski space-time is *less* absolute than some generic solution to the Einstein equations in the sense that the former has a large isometry group (viz. the Poincaré group, whose dimension is the maximal one an isometry group can have), whereas the isometry group of generic Lorentzian metrics is trivial. This also make GR *less* "covariant" than special relativity.

Against this, one may argue that in GR the metric is, after all, less absolute than in special relativity because in the full Einstein equations (1.22). it is coupled to the matter distribution. The point, then, is that there is no such coupling for (1.26), and this is supposed to make GR superior to special relativity because GR has no objects "that act but are not acted on". This fact is hidden by the vacuum field equation (1.25) and hence the argument rests on a distinction between the vacuum Einstein equations and those with matter. But this distinction is artificial. Furthermore, where does the buck stop? At least in its modern formulation GR uses smooth manifolds, which are modeled on $\mathbb{R}^4$ *with the usual smooth structure and topology*.[124] These are not dynamically generated but assumed, and hence should be counted as "absolute objects". Hence the argument, once it is carried through consistently, ultimately turns against GR, too.

The last argument we discuss is that of the particle physicist: much as the gauge invariance of electrodynamics expresses the fact that at the quantum level this theory describes interacting massless particles with helicity ± 1, i.e., *photons*, the general covariance of GR leads to massless particles with helicity ± 2, i.e., *gravitons* (see §8.5). This is arguably the strongest and physically most compelling argument for general covariance, but in the absence of even a perturbative theory of quantum gravity it is still feeble, not to speak of the wide gap between this kind of reasoning and the geometric structure of GR that leads to its general covariance in the first place.

Our conclusion is that while general covariance does not express a general relativity principle, it remains mysterious what it does express. Any resolution will have to navigate between:

- The pull towards physical relevance of general covariance in being a symmetry of the field equations of GR, whose ingredients (curvature and energy-momentum) are physical.[125]
- The pull against physical relevance, since no one has been able to figure it out so far.

At least in physical and mathematical practice, the second force has won: in its modern form of diffeomorphism invariance of Einstein's equations, general covariance is seen as a *gauge symmetry*, in the following sense.[126] Let M_i be a $4d$ manifold and g_i a Lorentzian metric on M_i solving the vacuum Einstein equations. Then two pairs (M_1, g_1) and (M_2, g_2) that differ by an isometry (i.e. a diffeomorphism that preserves the metric) describe the same space-time.[127]

This resolves the Hole Argument (at least in Hilbert's version, see Theorem 7.10) and also justifies the widely spread identification of active and passive coordinate transformations. Nonetheless, we will return to this discussion in the context of the "problem of time", which we regard as the "Hamiltonian shadow" of the problem of general covariance (see §8.11).

[124] Even with the usual topology there are innumerable inequivalent smooth structures on $\mathbb{R}^4$, each giving rise to a different concept of a smooth manifold. This is a result from *Donaldson theory* (Donaldson & Kronheimer, 1997).

[125] See e.g. Brading & Castellani (2003), Belot (2013), Caulton (2015), and Dewar (2019) for discussions of the concept of symmetry in physics. This concept is very tricky, as the debate on general covariance shows! Here we just call general coordinate transformations (or, for that matter, diffeomorphisms) symmetries because they are transformations that preserve solutions to the Einstein equations, cf. §1.5. *Defining* symmetries as transformations that preserve solutions is a 'recipe for disaster' (Belot, 2013, §3), but we use the idea the other way round.

[126] The gauge group depends on the details. In Einstein's GR it is a diffeomorphism group, but in other versions of GR it may consist of local Lorentz or Poincaré transformations (Blagojević & Hehl, 2013; Krasnov, 2020).

[127] But what does this mean? Suppose two stationary black holes have the same parameters and hence are both described by exactly the same Kerr metric, are they really "the same space-time"? The mind boggles!

2 General differential geometry

The mathematical language of GR is differential geometry, enriched by geometric analysis.

2.1 Manifolds

We start by reviewing the key definitions underlying the concept of a manifold.[128] First, a ***space*** means a topological space, assumed *Hausdorff*. The topology of M (i.e. the set of its open sets) is denoted by $\mathscr{O}(M)$, so that $U \in \mathscr{O}(M)$ means that $U \subset M$ and U is open. Since otherwise it cannot support a Lorentzian metric, in GR we may assume that M is also *metrizable*.[129] If M is a (topological) manifold, this is equivalent to M being *second countable* as well as *paracompact*.

Definition 2.1 1. *An n-dimensional* **(topological) manifold** *is a space M such that any $x \in M$ has a nbhd (= neighbourhood) $U \in \mathscr{O}(M)$ that is homeomorphic to some $V \in \mathscr{O}(\mathbb{R}^n)$. Equivalently, one may require V to be $\mathbb{R}^n$ itself, or some open ball in $\mathbb{R}^n$.*

2. *A* **chart** *on M is a pair (U,φ) where $U \in \mathscr{O}(M)$ and $\varphi : U \to \mathbb{R}^n$ is a homeomorphism onto its image $V = \varphi(U)$. A chart (U,φ) gives a* **coordinate system** *on U, in that the* **coordinates** *$(x^1,\dots,x^n)$ of $x \in U$ of x are $x^i = \varphi^i(x)$, where one writes $\varphi : U \to \mathbb{R}^n$ as $(\varphi^1,\dots,\varphi^n)$, where $\varphi^i : U \to \mathbb{R}$ in terms of the standard basis of $\mathbb{R}^n$ $(i = 1,\dots,n)$.*

3. *A C^k-***atlas** *on M (where $k \in \mathbb{N} \cup \{\infty\}$) is a collection of charts $(U_\alpha,\varphi_\alpha)$, where $M = \cup_\alpha U_\alpha$ (i.e. the U_α form an open cover of M), and, whenever $U_{\alpha\beta} = U_\alpha \cap U_\beta$ is not empty, writing $V_{\alpha\beta} = \varphi_\alpha(U_{\alpha\beta})$, the map $\varphi_\beta \circ \varphi_\alpha^{-1} : V_{\alpha\beta} \to \mathbb{R}^n$ is C^k (since $V_{\alpha\beta} \subset \mathbb{R}^n$ this is well defined).*

4. *Two C^k-atlases $(U_\alpha,\varphi_\alpha)$ and $(U'_{\alpha'},\varphi'_{\alpha'})$ on a topological manifold M are* **equivalent** *if their union is a C^k-atlas, i.e., if all transition functions $\varphi'_{\beta'} \circ \varphi_\alpha^{-1}$ and $\varphi_\beta \circ (\varphi'_{\alpha'})^{-1}$ (if defined) are C^k; this is indeed an equivalence relation. A* **C^k-structure** *on M is an equivalence class of C^k atlases on M. A* **C^k-manifold** *is a manifold with a C^k structure. A* **smooth manifold** *is a manifold with a C^∞ structure, that is, a C^∞-manifold.*

5. *A function $f : M \to \mathbb{R}$ on a smooth manifold is* **smooth**, *written $f \in C^\infty(M)$, if for some fixed atlas (within its equivalence class), each map $f \circ \varphi_\alpha^{-1} : V_\alpha \to \mathbb{R}$ is smooth.*[130]

6. *For two smooth manifolds M,N, a map $\psi : M \to N$ is* **smooth** *if for each $f \in C^\infty(N)$ the pullback $\psi^* f \equiv f \circ \psi$ is in $C^\infty(M)$. Equivalently, in terms of the manifolds: for any chart (U,φ) on M and chart $(\tilde{U},\tilde{\varphi})$ on N such that $U' = \psi(U) \cap \tilde{U} \neq \emptyset$, the function $\tilde{\varphi} \circ \psi \circ \varphi^{-1} : V' \to \tilde{V}$ is smooth (in the calculus sense), where $V' = \varphi(\psi^{-1}(U')) \subset V$.*

7. *A* **diffeomorphism** *of M is an invertible smooth map $\psi : M \to M$* with smooth inverse. *Under the obvious operations, such maps form the* **diffeomorphism group** *$\mathrm{Diff}(M)$ of M.*

Unless the contrary is stated, we henceforth assume that M is a smooth manifold equipped with some C^∞ atlas $(U_\alpha,\varphi_\alpha)$, and that all maps between smooth mathematical objects are smooth.

[128] See §2.6 for manifolds with boundary. References for this chapter are Choquet-Bruhat & DeWitt-Morette (1982), Abraham & Marsden (1985), Kriele (1999), Frankel (2004), Lee (2012), and Mărcuț (2016).

[129] See e.g. Palomo & Romero (2006), §1.1, or Minguzzi (2019), §1.8.

[130] This is then true for any atlas. Conversely, *M as a manifold* can be reconstructed from *$C^\infty(M)$ as a commutative algebra* via homomorphisms $\mathrm{ev} : C^\infty(M) \to \mathbb{R}$. See e.g. Navarro González & Sancho de Salas (2003), chapter 2.

2.2 Tangent bundle

The differential geometry relevant to GR comes from the tangent bundle, which generates the entire tensor calculus. Of the many roads to this bundle we prefer an (initially) algebraic construction in terms of derivations on $C^\infty(M)$, from which the geometric picture emerges.[131] Readers who are mainly interested in *using* tangent bundles can move straight to Definition 2.4.

Definition 2.2 *1. A* **derivation** *of an algebra A (over $\mathbb{R}$) is a linear map $\delta : A \to A$ satisfying*

$$\delta(ab) = \delta(a)b + a\delta(b). \tag{2.1}$$

2. For any smooth manifold M, a **point derivation** *at $x \in M$ is a linear map*

$$\delta_x : C^\infty(M) \to \mathbb{R} \tag{2.2}$$

that satisfies the **Leibniz rule**

$$\delta_x(fg) = \delta_x(f)g(x) + f(x)\delta_x(g). \tag{2.3}$$

In no. 2, $A = C^\infty(M)$ is seen as a (commutative) algebra with respect to pointwise operations.

The set $\mathrm{Der}(A)$ of all derivations of A is a vector space (again over $\mathbb{R}$). If A is associative and commutative, as is the case for $A = C^\infty(M)$, then $\mathrm{Der}(A)$ is also an A-module under the natural action $(a\delta)(b) = a\delta(b)$. In addition, $\mathrm{Der}(A)$ is a Lie algebra under the bracket

$$[\delta_1, \delta_2] := \delta_1 \circ \delta_2 - \delta_2 \circ \delta_1. \tag{2.4}$$

For $M = \mathbb{R}^n$, taking $X^i = \delta(x^i)$ it follows that each derivation δ of $C^\infty(\mathbb{R}^n)$ assumes the form

$$\delta(f)(x) = \sum_{j=1}^{n} X^j(x) \frac{\partial f(x)}{\partial x^j}, \tag{2.5}$$

where $X \in C^\infty(\mathbb{R}^n, \mathbb{R}^n)$, henceforth called $\mathfrak{X}(\mathbb{R}^n)$, is an "old-fashioned vector field" on $\mathbb{R}^n$, i.e. a field of arrows. Conversely, X defines a derivation $\delta \equiv \delta_X$ by reading (2.5) as a definition of δ. This gives a bijection $X \leftrightarrow \delta_X$ between the set $\mathfrak{X}(\mathbb{R}^n)$ of all *vector fields* on $\mathbb{R}^n$ and the set $\mathrm{Der}(C^\infty(\mathbb{R}^n))$ of all *derivations* on $C^\infty(\mathbb{R}^n)$. We further pass to point derivations by defining

$$\delta_x(f) := \delta(f)(x), \tag{2.6}$$

where $\delta \in \mathrm{Der}(C^\infty(\mathbb{R}^n))$. Conversely, Definition 2.2 implies that a family of point derivations $x \mapsto \delta_x$, defined for all $x \in \mathbb{R}^n$, comes from a single derivation δ via (2.6), and hence from a vector field X via $\delta = \delta_X$, iff the map $x \mapsto \delta_x(f)$ is smooth from $\mathbb{R}^n$ to $\mathbb{R}$ for each $f \in C^\infty(\mathbb{R}^n)$.

Eq. (2.4) also has a match for vector fields: $\mathfrak{X}(\mathbb{R}^n)$ is a Lie algebra under the ***commutator***

$$[X, Y](f) := X(Y(f)) - Y(X(f)). \tag{2.7}$$

[131] An ***algebra*** A (here always defined over $\mathbb{R}$) is a real vector space equipped with a bilinear map $A \times A \to A$, usually written $(a,b) \mapsto ab$. Many algebras are ***associative*** in that $(ab)c = a(bc)$ for all $a,b,c \in A$, as well as ***commutative***, i.e. $ab = ba$ for all $a,b \in A$. ***Lie algebras*** are neither: here one writes $(a,b) \mapsto [a,b]$, with axioms $[a,b] = -[b,a]$ as well as the ***Jacobi identity*** $[a,[b,c]] + [c,[a,b]] + [b,[c,a]] = 0$. A ***module*** over an algebra A is a vector space V with a bilinear map $A \times V \to V$, written $(a,v) \mapsto av$, such that $a(bv) = (ab)v$ (or $a(bv) = [a,b]v$).

In coordinates, where we use components $X = \sum_i X^i \partial_i$ and $Y = \sum_j Y^j \partial_j$, we have

$$[X,Y] = \sum_i [X,Y]^i \partial_i; \qquad [X,Y]^i = \sum_j (X^j \partial_j Y^i - Y^j \partial_j X^i). \tag{2.8}$$

Relative to (2.7) and (2.4), the bijection $X \leftrightarrow \delta_X$ is promoted to an isomorphism of Lie algebras.

Finally, if $\mathfrak{X}(\mathbb{R}^n)$ carries the $C^\infty(\mathbb{R}^n)$ action given by $(fX)^j(x) = f(x)X^j(x)$, then $X \leftrightarrow \delta_X$ is in addition an isomorphism of $C^\infty(\mathbb{R}^n)$ modules. Since $X : \mathbb{R}^n \to \mathbb{R}^n$ is given by its components $X^k : \mathbb{R}^n \to \mathbb{R}$, as a $C^\infty(\mathbb{R}^n)$ module $\mathfrak{X}(\mathbb{R}^n)$ decomposes as a direct sum of copies of $A = C^\infty(\mathbb{R}^n)$. By definition, this makes $\mathfrak{X}(\mathbb{R}^n)$ a ***free module*** over $C^\infty(\mathbb{R}^n)$. Of course, the same is then true for $\mathrm{Der}(C^\infty(\mathbb{R}^n))$. In sum, looking at a vector field X as the corresponding derivation δ_X, we often identify $\mathrm{Der}(C^\infty(\mathbb{R}^n))$ with $\mathfrak{X}(\mathbb{R}^n)$, and this identification preserves all relevant structure.

We now generalize this story to arbitrary manifolds M. On the algebraic side, we have the derivations $\mathrm{Der}(C^\infty(M))$. We are going to define vector fields geometrically as sections of the ***tangent bundle*** TM to M, whose construction is best understood in a more general form.

Definition 2.3 *A (real, locally trivial) k-dimensional* **vector bundle** *over M is an open surjective map $\pi : E \to M$, where E is a manifold, such that:*

1. *For each $x \in M$, the* **fiber** *$E_x := \pi^{-1}(x)$ is a k-dimensional (real) vector space, i.e. $E_x \cong \mathbb{R}^k$ (where k is independent of x). This is the main point. More technically:*
2. *M has an open cover (U_i) with diffeomorphisms $\Phi_i : \pi^{-1}(U_i) \to U_i \times \mathbb{R}^k$ such that:*
   (a) *Each restriction $\Phi_i : E_x \to \{x\} \times \mathbb{R}^k$ is an isomorphism of vector spaces ($x \in U_i$);*
   (b) *If $U_{ij} \equiv U_i \cap U_j \neq \emptyset$, then $\Phi_{ij} \equiv \Phi_i \circ \Phi_j^{-1} : U_{ij} \times \mathbb{R}^k \to U_{ij} \times \mathbb{R}^k$ is the identity on the first coordinate and a vector space isomorphism on the second one.*

A **vector bundle map** *from $\pi_1 : E \to M$ to $\pi_2 : F \to N$ is a* pair *of maps $\varphi_f : E \to F$ and $\varphi_b : M \to N$ such that $\pi_2 \circ \varphi_f = \varphi_b \circ \pi_1$, and each "fiber" map $\varphi_f : E_x \to F_{\varphi_b(x)}$ is linear.*

The simplest k-dimensional vector bundle over M is $E = M \times \mathbb{R}^k$ with π given by projection on the first coordinate; this is called a ***trivial bundle***. A ***(cross-)section*** of E is a map $s : M \to E$ such that $\pi \circ s = \mathrm{id}_M$, i.e., $\pi(s(x)) = x$ for each $x \in M$. The set of smooth sections of E is denoted by $\Gamma(E)$ or $\Gamma(M,E)$. This is a vector space. Under the natural action

$$C^\infty(M) \times \Gamma(E) \to \Gamma(E); \qquad (fs)(x) := f(x)s(x), \tag{2.9}$$

the vector space $\Gamma(E)$ is a finitely generated projective (f.g.p.) module over $C^\infty(M)$.[132]

Sections s of the trivial bundle $E = M \times \mathbb{R}^k \to M$ bijectively correspond to maps $\tilde{s} : M \to \mathbb{R}^k$ via $s(x) = (x, \tilde{s}(x))$. Hence we obtain, as an isomorphism of f.g.p. $C^\infty(M)$-modules,

$$\Gamma(M \times \mathbb{R}^k) \cong C^\infty(M, \mathbb{R}^k). \tag{2.10}$$

The ***Serre–Swan Theorem*** provides an isomorphism between f.g.p. modules $\mathscr{E}$ over $C^\infty(M)$ and vector bundles $E \to M$ over M, in such a way that $\mathscr{E} \cong \Gamma(E)$. We first define E as a set by

$$E := \sqcup_{x \in M} E_x; \qquad E_x := \mathscr{E} / \sim_x = \mathscr{E} / (C^\infty_x(M) \cdot \mathscr{E}). \tag{2.11}$$

[132] An A-module $\mathscr{E}$ is called ***finitely generated projective*** if there exists an A-module $\mathscr{F}$ such that $\mathscr{E} \oplus \mathscr{F}$ is free, i.e. isomorphic to a finite direct sum $\oplus^k A$. Equivalently, $\mathscr{E} \cong p(\oplus^k A)$ for some idempotent $p \in M_k(A)$ (i.e. $p^2 = p$).

I.e. $s_1 \sim_x s_2$ iff $s_1 - s_2 \in C^\infty_x(M)\cdot\mathscr{E}$, defined as the linear span in $\mathscr{E}$ of all fs, where $s\in\mathscr{E}$ and

$$f \in C^\infty_x(M) := \{f \in C^\infty(M) \mid f(x) = 0\}. \tag{2.12}$$

Then each fiber E_x of E is a vector space under the linear structure inherited from $\mathscr{E}$, that is,

$$\lambda[s_1]_x + \mu[s_2]_x := [\lambda s_1 + \mu s_2]_x; \qquad 0 := [0]_x, \tag{2.13}$$

where $[s]_x$ is the equivalence class of s with respect to $\sim_x$, and $\lambda,\mu\in\mathbb{R}$. Subsequently, define

$$\mathscr{E} \to \Gamma(E); \qquad \hat{s} \to s; \qquad s(x) = [\hat{s}]_x, \tag{2.14}$$

so that $s\in E_x$ and hence $s : M \to E$ is a cross-section of E. Then there is a unique smooth structure on E such that (2.14) is an isomorphism of $C^\infty(M)$ modules. This isomorphism maps $C^\infty_x(M)\cdot\mathscr{E}$ to $\Gamma(E;x) := \{s\in\Gamma(E) \mid s(x) = 0\}$, so that the mirror of (2.11) under the isomorphism (2.14) is

$$\Gamma(E)/\Gamma(E;x) \cong E_x. \tag{2.15}$$

We apply this to the $C^\infty(M)$-module $\mathscr{E} = \mathrm{Der}(C^\infty(M))$, and notice that we have an isomorphism

$$\mathrm{Der}(C^\infty(M))/\sim_x \overset{\cong}{\to} \mathrm{Der}_x(C^\infty(M)); \qquad [\delta]_x \mapsto \delta_x, \tag{2.16}$$

where $\mathrm{Der}_x(C^\infty(M))$ is the vector space of all point derivations δ_x of M, cf. (2.2) - (2.3). Although $\mathrm{Der}(C^\infty(M))$ may no longer be free (as in $M=\mathbb{R}^n$), using charts one can show that it is finitely generated projective, so that the above procedure for defining a vector bundle E is applicable.

Definition 2.4 *The* **tangent bundle** $\pi : TM \to M$ *is the vector bundle E constructed from*

$$\mathscr{E} = \mathrm{Der}(C^\infty(M)) \tag{2.17}$$

as in the above procedure, replacing (2.11) *by* (2.16)*. That is, the total space and fibers are*

$$TM := \sqcup_{x\in M} T_xM; \qquad T_xM := \mathrm{Der}_x(C^\infty(M)), \tag{2.18}$$

and the smooth structure of TM is (uniquely) defined by the property that the map

$$\mathrm{Der}(C^\infty(M)) \to \mathfrak{X}(M) := \Gamma(TM); \qquad \delta \mapsto (x \mapsto \delta_x), \tag{2.19}$$

where δ_x is defined by (2.6)*, is an isomorphism. A* **vector field** *on M is a cross-section of TM.*

In a local chart $\varphi : U \to \mathbb{R}^n$, for $x\in U$ we define the symbol ∂_i as an element of T_xM by

$$\partial_i f(x) := \frac{\partial(f\circ\varphi^{-1})}{\partial x^i}(\varphi(x)), \tag{2.20}$$

where $f\in C^\infty(U)$ and φ^{-1} is the inverse of $\varphi : U \to V = \varphi(U)$. With $V\subset\mathbb{R}^n$, the function $f\circ\varphi^{-1} : V\to\mathbb{R}$ is the coordinate expression $f(x^1,\dots,x^n)$ of f, so that ∂_i in (2.20) may be taken literally. This also shows that $(\partial_i,\dots,\partial_n)$ is a basis of T_xM, so that we may expand $X_x\in T_xM$ as

$$X_x = \sum_{i=1}^n X^i_x \partial_i; \qquad X^i_x = X\varphi^i(x), \tag{2.21}$$

where $\varphi = (\varphi^1,\dots,\varphi^n) : U\to\mathbb{R}^n$. Thus TM is an n-dimensional vector bundle over M.

In conclusion, a *vector field* on M, written $X \in \mathfrak{X}(M)$, is a map $x \mapsto X_x$, also written as $x \mapsto X(x)$, where $x \in M$ and $X_x \in T_xM$ as defined by (2.18). A *derivation* on M is a map $\delta : C^\infty(M) \to C^\infty(M)$ that satisfies (2.1). These concepts are related by (2.6) with $\delta_x = X_x$. We think of a vector field $X \in \mathfrak{X}(M)$ as the *collection* of all "tangent vectors" $X_x \in T_xM$, whereas we think of the corresponding derivation δ as a *single* global operation on $C^\infty(M)$.

- *Point* derivations ***push forward*** under maps $\psi : M \to N$: for $x \in M$ we have linear maps

$$T_x\psi \equiv \psi'_x : T_xM \to T_{\psi(x)}N; \qquad (\psi'_x\delta_x)(g) = \delta_x(\psi^* g) \ \ (g \in C^\infty(N)), \tag{2.22}$$

  where $\psi^* g := g \circ \psi$ is the ***pullback*** of g. Collecting these maps gives a vector bundle map

$$T\psi \equiv \psi_* \equiv \psi' : TM \to TN. \tag{2.23}$$

- However, *derivations* (or vector fields) push forward only if $\psi : M \to N$ is a *diffeomorphism*: the map $\psi_* : \mathrm{Der}(C^\infty(M)) \to \mathrm{Der}(C^\infty(N))$, or $\psi_* : \mathfrak{X}(M) \to \mathfrak{X}(N)$, is given by

$$\psi_*(\delta) = (\psi^{-1})^* \circ \delta \circ \psi^*. \tag{2.24}$$

  One needs $(\psi^{-1})^*$ even if $N = M$, since $\delta \circ \psi^*$ fails to be a derivation of $C^\infty(M)$. Check!

So far, tangent vectors $X_x \in T_xM$ were defined *algebraically* as point derivations, i.e. as linear maps $\delta_x : C^\infty(M) \to \mathbb{R}$ satisfying (2.3). *Geometrically*, each tangent vector (*nomen est omen!*) is tangent to some ***curve*** γ through x, i.e., a map $\gamma : I \to M$, where $I \subset \mathbb{R}$ is some interval we always assume to contain 0, such that $\gamma(0) = x$ (see below for the existence of γ). In other words,

$$X_x(f) = \frac{d}{dt} f(\gamma(t))_{|t=0}, \tag{2.25}$$

which *symbolically* may be written as $X_x = \dot\gamma \equiv d\gamma/dt$, or even as $X_x = d/dt$, with γ understood. This description gives a geometric perspective on the pushforward of T_xM just described:

- If $X = d\gamma/dt$ is tangent to γ, then $\psi' X = d(\psi \circ \gamma)/dt$ is tangent to $\psi(\gamma)$.

In a chart $\varphi = (\varphi^1, \dots, \varphi^n) : U \to \mathbb{R}^n$, with $x \in U$, the components X^i_φ of X_x are given by

$$X^i_\varphi = X\varphi^i(x) = \frac{d}{dt}\varphi^i(\gamma(t))_{|t=0} = \frac{d}{dt}\gamma^i(t)_{|t=0}, \tag{2.26}$$

where $\gamma^i(t) = \varphi^i(\gamma(t))$. This also shows that γ exists, given X_x, since it just has to satisfy (2.26). Of course, γ is far from unique. Eq. (2.26) gives the traditional transformation rule for vectors under a change of charts (i.e. of coordinates). If $x \in U_\alpha \cap U_\beta$, then (2.25) and (2.26) imply

$$X^i_\beta = \sum_j \frac{\partial x^i_\beta}{\partial x^j_\alpha} X^j_\alpha, \tag{2.27}$$

where $X^i_\beta \equiv X^i_{\varphi_\beta}$ etc., and each coordinate $x^i_\beta = \varphi^i_\beta(x)$ of x with respect to φ_β is seen as a function of all coordinates $x^i_\alpha = \varphi_\alpha(x)$ of x with respect to φ_α via the identity $\varphi^i_\beta = \varphi^i_\beta \circ \varphi^{-1}_\alpha \circ \varphi_\alpha$, i.e.

$$x^i_\beta(x_\alpha) = \varphi^i_\beta \circ \varphi^{-1}_\alpha(x_\alpha). \tag{2.28}$$

It is important to distinguish (2.27), which is a ***change of coordinates formula*** for a given tangent vector, from the ***pushforward of a tangent vector*** under a map $\psi : M \to M$. With $\varphi : U \to V \subset \mathbb{R}^n$, suppose for simplicity that $x \in U$ and also $\psi(x) \in U$. Then, writing $X^i_\varphi \equiv X^i$ as above, as well as $\psi^i = \varphi^i \circ \psi \circ \varphi^{-1}$ (which near x is a function from V to $\mathbb{R}$), we have

$$(\psi' X)^i = \sum_j \frac{\partial \psi^i}{\partial x^j} X^j. \tag{2.29}$$

A curve $\gamma : I \to M$ ***integrates*** a vector *field* X if $X_{\gamma(t)} = d\gamma(t)/dt$ for all $t \in I$, i.e., in coordinates,

$$\frac{d\gamma^j(t)}{dt} = X^j(\gamma^1(t), \ldots, \gamma^n(t)), \ (j = 1, \ldots, n). \tag{2.30}$$

The theory of ODEs shows that for each $x \in M$ there exists an open interval $I \subset \mathbb{R}$ (with $0 \in I$) and a curve $\gamma : I \to M$ on which (2.30) holds with $\gamma(0) = x$. This solution is unique in the sense that if $\gamma_1 : I_1 \to M$ and $\gamma_2 : I_2 \to M$ both satisfy (2.30) with $\gamma_1(0) = \gamma_2(0) = x_0$, then $\gamma_1 = \gamma_2$ on $I_1 \cap I_2$. Taking unions, it follows that there exists a maximal interval I on which γ is defined.

If for any $x \in M$ there is a curve $\gamma : \mathbb{R} \to M$ satisfying (2.30) with $\gamma(0) = x$, we say that $X \in \mathfrak{X}(M)$ is ***complete***.[133] In that case, all integrating curves γ can be assembled into the ***flow*** of X. This is a smooth map $\psi : \mathbb{R} \times M \to M$, written $\psi_t(x) \equiv \psi(t,x)$, that satisfies

$$\psi_0(x) = x; \tag{2.31}$$

$$X_{\psi_t(x)} f = \frac{d}{dt} f(\psi_t(x)) \tag{2.32}$$

for all $x \in M$, $t \in \mathbb{R}$, and $f \in C^\infty(M)$. Thus the flow ψ of X gives "the" integral curve γ of X through x_0 by $\gamma(t) = \psi_t(x_0)$. Any complete vector field has a unique flow. Uniqueness implies that M is a disjoint union of the integral curves of X (which can never cross each other because of the uniqueness of the solution), and also implies the composition rule

$$\psi_s \circ \psi_t = \psi_{s+t}. \tag{2.33}$$

From a group-theoretic point of view, a flow is therefore an action of $\mathbb{R}$ (as an additive group) on M *that in addition integrates* X in the sense of (2.32). In particular, (2.33) implies $\psi_{-t} = \psi_t^{-1}$, so that each $\psi_t : M \to M$ is automatically a diffeomorphism of M.

If X is not complete (a case that will be of great interest to GR!), we first define the ***domain*** $D_X \subset \mathbb{R} \times M$ of ψ as the set of all $(t,x) \in \mathbb{R} \times M$ for which there exists an open interval $I \subset \mathbb{R}$ containing 0 and t, as well as a (necessarily unique) curve $\gamma : I \to M$ that satisfies (2.30) with initial condition $\gamma(0) = x$. Obviously $\{0\} \times M \subset D_X$, and (less trivially) it turns out that D_X is open. Then a flow of X is a map $\psi : D_X \to M$ that satisfies (2.31) for all x and (2.32) for all $(t,x) \in D_X$. Eq. (2.33) then holds if the left-hand side (and hence the right-hand side) is defined.

As a first application of flows, let us define the ***Lie derivative*** $\mathscr{L}_X Y$ of some vector field $Y \in \mathfrak{X}(M)$ with respect to another vector field $X \in \mathfrak{X}(M)$ by

$$\mathscr{L}_X Y(x) = \lim_{t \to 0} \frac{Y_{\psi_t(x)} - \psi'_t(Y_x)}{t} = \lim_{t \to 0} \frac{\psi'_{-t}(Y_{\psi_t(x)}) - Y_x}{t} \tag{2.34}$$

where ψ is the flow of X. Note that $Y_{\psi_t(x)} - Y_x$ would be undefined, since $Y_{\psi_t(x)} \in T_{\psi_t(x)}M$ whilst $Y_x \in T_x M$ and these are different vector spaces; the pushforward ψ'_t serves to move Y_x to $T_{\psi_t(x)}M$. A simple computation then yields the extremely useful result

$$\mathscr{L}_X Y = [X, Y]. \tag{2.35}$$

[133] If X has *compact support*, then it is complete. So if M is compact, then every vector field on M is complete.

2.3 Dual vector spaces, metrics, and tensor products

In order to defined tensors we need some linear algebra. Let V be a finite-dimensional real vector space, with $\dim(V) = n$, which in GR will be $V = T_xM$. The ***dual*** $V^* = \mathrm{Hom}(V, \mathbb{R})$ consists of all linear maps from V to $\mathbb{R}$. This is a real vector space in its own right under pointwise constructions. It is isomorphic to V (as a vector space), but not canonically so: one needs to specify a basis $(e_1, \ldots, e_n)$ of V, with corresponding dual basis $(\omega^1, \ldots, \omega^n)$ defined by $\omega^a(e_b) = \delta^a_b$, upon which the ugly map $\sum_a v^a e_a \mapsto \sum_a v^a \omega^a$ from V to V^* is an isomorphism (which obviously depends on the chosen basis). However, we do have a canonical isomorphism

$$V \cong V^{**}; \qquad v \mapsto \hat{v}; \qquad \hat{v}(\theta) = \theta(v) \tag{2.36}$$

where $\hat{v} \in V^{**} = \mathrm{Hom}(V^*, \mathbb{R})$. This map is injective for any V, but it is surjective (and hence an isomorphism) iff V is finite-dimensional. One often writes $\langle \theta, v \rangle$ for both $\theta(v)$ and $\hat{v}(\theta)$.

The naturality of the isomorphism $V^* \cong V$ improves markedly in the presence of a *metric*.

Definition 2.5 *A* **metric** *g on V is a bilinear map $g : V \times V \to \mathbb{R}$ that is:*

- symmetric, *in that $g(v,w) = g(w,v)$ for all $v, w \in V$;*
- nondegenerate, *i.e. for each nonzero vector $v \in V$ there is $w \in V$ such that $g(v,w) \neq 0$.*

A metric g yields two maps that are mutually inverse and hence are isomorphisms $V^* \cong V$:

$$\flat : V \to V^*, \qquad \flat(v) \equiv v^\flat; \qquad v^\flat(w) := g(v,w); \tag{2.37}$$

$$\sharp : V^* \to V, \qquad \sharp(\theta) \equiv \theta_\sharp; \qquad g(\theta_\sharp, v) := \theta(v). \tag{2.38}$$

Any metric g can be diagonalized, i.e. V has an ***orthonormal*** basis $(e_a) \equiv (e_1, \ldots, e_n)$, in which

$$g(e_a, e_b) = \varepsilon_a \delta_{ab}; \qquad \varepsilon_a = \pm 1. \tag{2.39}$$

The pair (n_-, n_+), where n_-/n_+ is the number of negative/positive numbers ε_a, is independent of the basis and hence is an intrinsic property of a metric g, called its ***signature***. Especially in relativity, the signature is often written as $(- \cdots - + \cdots +)$, with n_-/n_+ minus/plus signs.

We now turn to the tensor product. In the following proposition, V and W are real but not necessarily finite-dimensional (and the same construction works over any field, typically $\mathbb{C}$).

Proposition 2.6 *Let V and W be real vector spaces. There is a real vector space called $V \otimes W$, in words the* **tensor product** *of V and W (over $\mathbb{R}$), and a map*

$$p : V \times W \to V \otimes W; \qquad p(v,w) \equiv v \otimes w, \tag{2.40}$$

such that for any vector space X and any bilinear map $\beta : V \times W \to X$, there is a unique linear map $\beta' : V \otimes W \to X$ such that $\beta = \beta' \circ p$. In other words, the following diagram commutes:

$$\begin{array}{ccc} V \times W & \xrightarrow{\ p\ } & V \otimes W \\ & \searrow^{\beta} & \downarrow{\scriptstyle \exists! \beta'} \\ & & X \end{array} \tag{2.41}$$

Moreover, this so-called universal property implies that $V \otimes W$ is unique up to isomorphism.

We will not prove this here in general, but do show existence of $V \otimes W$ if V and W are finite-dimensional.[134] We first assume that $V = Y^*$ and $W = Z^*$, in which case we define

$$Y^* \otimes Z^* := \mathrm{Hom}(Y \times Z, \mathbb{R}); \tag{2.42}$$

$$(\sigma \otimes \rho)(y,z) := \sigma(y)\rho(z), \tag{2.43}$$

where $\mathrm{Hom}(Y \times Z, \mathbb{R})$ is the space of bilinear maps from $Y \times Z$ to $\mathbb{R}$, and of course $\sigma \in Y^*$, $\rho \in Z^*$, $y \in Y$, $z \in Z$. Then $\beta'(\sigma \otimes \rho) = \beta(\sigma, \rho)$ by construction, and this uniquely extends to a linear map $\beta' : \mathrm{Hom}(Y \times Z, \mathbb{R}) \to X$, since $\mathrm{Hom}(Y \times Z, \mathbb{R})$ is the linear span of all $\sigma \otimes \rho$.

This also covers V and W themselves, at least up to isomorphism, since in finite dimension we have the isomorphism (2.36), so that, identifying V with V^{**} etc., we obtain

$$V \otimes W \cong V^{**} \otimes W^{**} = \mathrm{Hom}(V^* \times W^*, \mathbb{R}); \tag{2.44}$$

$$(v \otimes w)(\theta, \tau) = \theta(v)\tau(w), \tag{2.45}$$

where this time $v \in V$, $w \in W$, $\theta \in V^*$, and $\tau \in W^*$. Once again, $\beta' : \mathrm{Hom}(V^* \times W^*, \mathbb{R}) \to X$ is uniquely defined by linear extension of $\beta'(v \otimes w) = \beta(v, w)$, since the linear span of all $v \otimes w$ equals $\mathrm{Hom}(V^* \times W^*, \mathbb{R})$. We have effectively identified v with $\hat{v}$ and w with $\hat{w}$, cf. (2.36), and this shows up: although (2.42) gives $Y^* \otimes \mathbb{R} = Y^*$ as expected, eq. (2.44) has the consequence

$$V \otimes \mathbb{R} = \mathrm{Hom}(V^*, \mathbb{R}) = V^{**}, \tag{2.46}$$

where one would prefer to see V. But although no one would criticize the realization $V \otimes \mathbb{R} = V$, eq. (2.46) reconfirms that tensor products are merely defined *up to isomorphism*, cf. (2.36). Similarly, instead of $V^* \otimes V = \mathrm{Hom}(V \times V^*, \mathbb{R})$, as suggested by (2.42) and (2.44), we may take

$$V^* \otimes V \cong \mathrm{Hom}(V, V), \tag{2.47}$$

since one has an isomorphism $\mathrm{Hom}(V \times V^*, \mathbb{R}) \to \mathrm{Hom}(V, V)$, given by linear extension of

$$w \otimes \theta \mapsto (v \mapsto \theta(v)w). \tag{2.48}$$

With $v \in V$ and $\theta \in V^*$ as before, the inverse of the map (2.48) is given by

$$\mathrm{Hom}(V, V) \to \mathrm{Hom}(V \times V^*, \mathbb{R}); \qquad \varphi \mapsto \hat{\varphi}; \qquad \hat{\varphi}(v, \theta) = \theta(\varphi(v)). \tag{2.49}$$

In connection with the Riemann tensor we will have occasion to use the induced isomorphism

$$V^* \otimes V^* \otimes W^* \otimes W \cong \mathrm{Hom}(V \times V, \mathrm{Hom}(W, W)); \tag{2.50}$$

$$\theta_1 \otimes \theta_2 \otimes \eta \otimes w_1 \mapsto ((v_1, v_2) \mapsto (w_2 \mapsto \theta_1(v_1)\theta_2(v_2)\eta(w_2)w_1)). \tag{2.51}$$

To describe the inverse of this map we combine (2.42) and (2.44) to pick the realization

$$V^* \otimes V^* \otimes W^* \otimes W = \mathrm{Hom}(V \times V \times W \times W^*, \mathbb{R}). \tag{2.52}$$

The image $\hat{\varphi} \in \mathrm{Hom}(V \times V \times W \times W^*, \mathbb{R})$ of $\varphi \in \mathrm{Hom}(V \otimes V, \mathrm{Hom}(W, W))$ is then given by

$$\hat{\varphi}(v_1, v_2, w, \eta) = \eta(\varphi(v_1, v_2)(w)). \tag{2.53}$$

[134] The construction applies in general if we define $V \otimes W$ as the finite linear span of all $a \otimes b$ in $\mathrm{Hom}(V^* \times W^*, \mathbb{R})$.

2.4 Cotangent bundle

Now that we have the tangent bundle TM and the constructions in §2.3, all relevant vector bundles on M that are relevant for GR follow. First, the ***cotangent bundle*** T^*M is defined as

$$T^*M := \sqcup_{x\in M} T_x^*M; \qquad T_x^*M \equiv (T_xM)^* := \mathrm{Hom}(T_xM, \mathbb{R}), \tag{2.54}$$

i.e. T_xM is the dual of the vector space T_xM, consisting of all linear maps $\theta_x : T_xM \to \mathbb{R}$. The smooth structure of T^*M is the unique one such that elements $\theta \in \Gamma(T^*M) \equiv \Omega^1(M) \equiv \Omega(M)$, called ***covectors*** (or ***1-forms***), consist of those maps $x \mapsto \theta_x$ for which the function $x \mapsto \theta_x(X_x)$ from M to $\mathbb{R}$ is smooth for each vector field $X \in \mathfrak{X}(M)$. Since $T_xM \cong \mathbb{R}^n$ we also have $T_x^*M \cong \mathbb{R}^n$, so that, like the tangent bundle TM, also the cotangent bundle T^*M is an n-dimensional vector bundle over M. In a coordinate systems (x^i) defined by some chart, T_x^*M has basis $(dx^1, \ldots, dx^n)$ defined by $dx^i(\partial_j) = \delta^i_j$, which is dual to the basis $(\partial_1, \ldots, \partial_n)$ of T_xM defined in (2.20). Thus

$$\theta = \sum_i \theta_i dx^i; \qquad \theta_i = \theta(\partial_i). \tag{2.55}$$

For an equivalent view of dx^i, one may define the ***exterior derivative*** $d : C^\infty(M) \to \Omega(M)$ by

$$df(X) := X(f). \tag{2.56}$$

Then dx^i coincides with $d\varphi^i$, where $x^i = \varphi^i(x)$ as usual, and in coordinates (2.56) simply reads

$$df = \sum_i \left(\frac{\partial f}{\partial x^i}\right) dx^i. \tag{2.57}$$

More generally, let (e_a) be a basis of T_xM, with dual basis (ω^a) of T_x^*M (i.e. $\omega^a(e_b) = \delta^a_b$). Once again, if we expand $\theta = \sum_a \theta_a \omega^a$, we have $\theta_a = \theta(e_a)$. This may be done at a single point, but bases like $(\partial_1, \ldots, \partial_n)$ and $(dx^1, \ldots, dx^n)$ are defined at each $x \in U$ on which the coordinates $x^i = \varphi^i(x)$ are defined. Similarly, some basis (e_a) may be defined at each $x \in U$, where $U \in \mathcal{O}(M)$ need not even be the domain of a chart. In that case (e_a) is called a (moving) ***frame*** or an ***n-bein*** (so that in GR one has a ***vierbein*** or ***tetrad***). Abstractly, if $E \to M$ is a k-dimensional vector bundle, one may locally find k linearly independent cross-sections $(u_1, \ldots, u_k)$ of E and expand any $s \in \Gamma(E)$ by $s(x) = \sum_j s_j(x) u_j(x)$, where $s_j \in C^\infty(M)$ and $u_j \in \Gamma(E)$.

Whereas tangent vectors *push forward* from M to N under maps $\psi : M \to N$, covectors ***pull back*** from N to M, like functions: besides the pull-back $\psi^* : C^\infty(N) \to C^\infty(M)$ on functions, any (smooth) map ψ induces a pullback $\psi^* : \Omega(N) \to \Omega(M)$ on 1-forms by

$$(\psi^*\theta)_x(X_x) = \theta_{\psi(x)}(\psi'_x X_x), \tag{2.58}$$

where $\theta \in \Omega(N)$ and $X_x \in T_xM$. For any $f \in C^\infty(N)$ with $df \in \Omega(N)$, this yields

$$\psi^*(df) = d(\psi^* f). \tag{2.59}$$

A decent vector bundle map $\psi^* : T^*N \to T^*M$ is defined only if ψ is a diffeomorphism: for $\theta_y \in T_y^*N$ ($y \in N$), we need $x = \psi^{-1}(y) \in M$, so that the pullback $\psi_y^*(\theta_y) \in T_x^*M$ is defined by

$$(\psi_y^*\theta_y)(X_x) = \theta_y(\psi'_x X_x). \tag{2.60}$$

If ψ is merely injective, then we still obtain a map $\psi^* : T^*(\psi(M)) \to T^*M$ in this way.

2.5 Tensor bundles

For $(k,l) \in \mathbb{N} \times \mathbb{N}$ we define a vector bundle $T^{(k,l)}M$ over M in the usual way via its fibers

$$T_x^{(k,l)}M := \mathrm{Hom}((T_xM)^k \times (T_x^*M)^l, \mathbb{R}), \tag{2.61}$$

i.e. the vector space of $(k+l)$-fold multilinear maps from $(T_xM)^k \times (T_x^*M)^l$ to $\mathbb{R}$. These fibers comprise the total space of the bundle as a disjoint union

$$T^{(k,l)}M := \sqcup_{x\in M} T_x^{(k,l)}M, \tag{2.62}$$

whose manifold structure will be defined below (by defining the smooth sections). We then have

$$T^{(0,0)}M = M \times \mathbb{R}; \qquad T^{(1,0)}M = T^*M; \qquad T^{(0,1)}M \cong TM, \tag{2.63}$$

where in the last entry we used (2.36). Repeatedly using Proposition 2.6, taking (2.44) as a realization of "the" tensor product, and once again using (2.36), we obtain the realization

$$T_x^{(k,l)}M \cong (\otimes^k T_x^*M) \otimes (\otimes^l T_xM), \tag{2.64}$$

where $\otimes^l V$ is the l times iterated tensor product of V with itself. According to (2.64), the fiber $T_x^{(k,l)}M$ consists of finite sums of elementary tensors $\alpha_1 \otimes \cdots \otimes \alpha_k \otimes v_1 \otimes \cdots \otimes v_l$, defined for

$$\alpha_i \in T_x^*M (i = 1,\ldots,k); \qquad v_j \in T_xM (j = 1,\ldots,l).$$

In terms of (2.61), one has

$$\alpha_1 \otimes \cdots \otimes \alpha_k \otimes v_1 \otimes \cdots \otimes v_l(X_1,\ldots,X_k;\theta^1,\ldots,\theta^l) = \alpha_1(X_1)\cdots\alpha_k(X_k)v_1(\theta^1)\cdots v_l(\theta^l),$$

where each $X_i \in T_xM$ and each $\theta^j \in T_x^*M$. We then define $\Gamma(T^{(k,l)}M)$ as the set of all cross-sections $x \mapsto \tau_x$ from M to $T^{(k,l)}M$ (i.e. maps such that $\tau_x \in T_x^{(k,l)}M$) for which the map

$$x \mapsto \tau_x(X_1(x),\ldots,X_k(x);\theta^1(x),\ldots,\theta^l(x))$$

from M to $\mathbb{R}$ is smooth for each $(X_1,\ldots,X_k;\theta^1,\ldots,\theta^l)$ with $X_i \in \mathfrak{X}(M)$ and $\theta^j \in \Omega(M)$. This equips the vector bundles $T^{(k,l)}M$ with a manifold structure, in that we declare $\Gamma(T^{(k,l)}M)$ to be the space of smooth cross-sections of $T^{(k,l)}M$. Elements of $\Gamma(T^{(k,l)}M)$ are called ***tensors*** (or ***tensor fields*** (if τ_x is regarded as a tensor). In GR, $T^{(2,0)}M$ and $T^{(3,1)}M$ will be very important.

All this can be rewritten in terms of ***indices***. In terms of the (coordinate) basis $(\partial_1,\ldots,\partial_n)$ of T_xM with dual basis $(dx^1,\ldots,dx^n)$ of T_x^*M, the fiber $T_x^{(k,l)}M$ then has a basis

$$(dx^{i_1} \otimes \cdots \otimes dx^{i_k} \otimes \partial_{j_1} \otimes \cdots \otimes \partial_{j_l}), \tag{2.65}$$

where all indices run from 1 to n. *Thus $T^{(k,l)}M$ is an n^{k+l}-dimensional vector bundle*. Like vectors, tensors at x may be specified by their components with respect to some basis of T_xM and associated dual basis of T_x^*M. In the usual coordinate basis (∂_i) we have

$$\tau_x = \tau_{i_1\cdots i_k}^{j_1\cdots j_l}(x)\, dx^{i_1} \otimes \cdots dx^{i_k} \otimes \partial_{j_1} \otimes \cdots \otimes \partial_{j_l}; \tag{2.66}$$

$$\tau_{i_1\cdots i_k}^{j_1\cdots j_l}(x) = \tau_x(\partial_{i_1},\ldots,\partial_{i_k};dx^{j_1},\ldots,dx^{j_l}), \tag{2.67}$$

where we use the ***Einstein summation convention: repeated indices are summed over***.

That is, the right-hand side of (2.66) should really be preceded by $\sum_{i_1,\dots,i_k,j_1,\dots,j_l=1}^{n}$.

Similarly, in an arbitrary basis (e_a) of T_xM with dual basis (θ^a) of T_x^*M one has

$$\tau_x = \tau^{b_1\cdots b_l}_{a_1\cdots a_k}(x)\,\theta^{a_1}\otimes\cdots\theta^{a_k}\otimes e_{b_1}\otimes\cdots\otimes e_{b_l}; \tag{2.68}$$

$$\tau^{b_1\cdots b_l}_{a_1\cdots a_k}(x) = \tau_x(e_{a_1},\dots,e_{a_k};\theta^{b_1},\dots,\theta^{b_l}). \tag{2.69}$$

We write $\mathfrak{X}^{(k,l)}(M)$ for the space of cross-sections $\Gamma(T^{(k,l)}M)$ of $T^{(k,l)}M$, so that

$$\mathfrak{X}^{(0,0)}(M) = C^\infty(M); \qquad \mathfrak{X}^{(0,1)}(M) = \mathfrak{X}(M); \qquad \mathfrak{X}^{(1,0)}(M) = \Omega(M). \tag{2.70}$$

A tensor $\tau\in\mathfrak{X}^{(k,l)}(M)$ of type (k,l) maps k vector fields $(X_1,\dots,X_k)$ and l covector fields $(\theta^1,\dots,\theta^l)$ to a smooth function on M by pointwise evaluation, i.e.

$$\tau : \mathfrak{X}(M)^k\times\Omega(M)^l\to C^\infty(M); \tag{2.71}$$

$$\tau(X_1,\dots,X_k,\theta^1,\dots,\theta^l) : x\mapsto \tau_x(X_1(x),\dots.X_k(x);\theta^1(x),\dots,\theta^l(x)). \tag{2.72}$$

This map is evidently $k+l$-multilinear linear over $C^\infty(M)$, in the sense that e.g.

$$\tau(f_1X_1,\dots,f_kX_k,g_1\theta^1,\dots,g_l\theta^l) = f_1\cdots f_k\cdot g_1\cdots g_l\cdot\tau(X_1,\dots,X_k;\theta^1,\dots,\theta^l), \tag{2.73}$$

for all $f_i,g_j\in C^\infty(M)$; here we use the fact that $\mathfrak{X}(M)$ and $\Omega(M)$ are $C^\infty(M)$ modules.

Proposition 2.7 **(tensoriality test)** *A map*

$$\tau : \mathfrak{X}(M)^k\times\Omega(M)^l\to C^\infty(M) \tag{2.74}$$

is given by a tensor

$$\tau\in\mathfrak{X}^{(k,l)}(M) \tag{2.75}$$

through (2.72) *iff* τ *satisfies* (2.73)*, i.e., iff it is* $C^\infty(M)$*-multilinear in all entries.*

Proof. The proof is easy in local coordinates, where (2.73) yields

$$\begin{aligned}\tau(X_1,\dots,X_k,\theta^1,\dots,\theta^l) &= \tau(X_1^{i_1}\partial_{i_1},\dots,X_k^{i_k}\partial_{i_k};\theta^1_{j_1}dx^{j_1},\dots\theta^l_{j_l}dx^{j_l})\\ &= X_1^{i_1}\cdots X_k^{i_k}\cdot\theta^1_{j_1}\cdots\theta^l_{j_l}\tau(\partial_{i_1},\dots,\partial_{i_k};dx^{j_1},\dots dx^{j_l}),\end{aligned} \tag{2.76}$$

so if we define the components $\tau^{j_1\cdots j_l}_{i_1\cdots i_k}(x)$ of τ_x by (2.67) and subsequently define τ_x itself by (2.66), we have found the desired tensor that via (2.72) reproduces the given map τ. □

Eqs. (2.66) - (2.67) imply the transformation properties of tensors under changes of coordinates (i.e. charts), which historically even *defined* tensors: in the situation of (2.27),

$$(\tau_\beta)^{j_1\cdots j_l}_{i_1\cdots i_k}(x_\beta) = \frac{\partial x^{j_1}_\beta}{\partial x^{j'_1}_\alpha}\cdots\frac{\partial x^{j_l}_\beta}{\partial x^{j'_l}_\alpha}\cdot\frac{\partial x^{i'_1}_\alpha}{\partial x^{i_1}_\beta}\cdots\frac{\partial x^{i'_k}_\alpha}{\partial x^{i_k}_\beta}\cdot(\tau_\alpha)^{j'_1\cdots j'_l}_{i'_1\cdots i'_k}(x_\alpha), \tag{2.77}$$

where the "new" coordinates $(x_\beta)=(x^1_\beta,\dots,x^n_\beta)$ are functions of the "old" coordinates $(x_\alpha)=(x^1_\alpha,\dots,x^n_\alpha)$, cf. (2.28), and hence the matrix $(\partial x^{i'_1}_\alpha/\partial x^{i_1}_\beta)$ is defined as the inverse of the matrix $(\partial x^{i_1}_\beta/\partial x^{i'_1}_\alpha)$, *both* seen as functions of the (x^i_α). Note that the argument x_β in (2.77) refers to the

same point $x \in M$ as the argument x_α (but in *different coordinates*). Conversely, from a "tensor" (in the original historical sense of the word) $\tau^{j_1\cdots j_l}_{i_1\cdots i_k}(x)$ we obtain a map τ of the kind (2.71) by

$$\tau(X_1,\ldots,X_k,\theta^1,\ldots,\theta^l) : x \mapsto \ \tau^{j_1\cdots j_l}_{i_1\cdots i_k}(x) X_1^{i_1}(x)\cdots X_k^{i_k}(x)\cdot\theta^1_{j_1}(x)\cdots\theta^l_{j_l}(x). \tag{2.78}$$

It then follows from (2.77) that τ is well defined in being coordinate-independent. It is also $k+l$-multilinear linear over $C^\infty(M)$ by construction, so that we recover (2.71) - (2.73).

A smooth map

$$\psi : M \to N \tag{2.79}$$

induces a (vector bundle) map

$$\psi_*^{(0,l)} : T^{(0,l)}M \to T^{(0,l)}N \tag{2.80}$$

via the obvious pointwise maps

$$\psi_x^{(0,l)} : T_x^{(0,l)}M \to T^{(0,l)}_{\psi(x)}N. \tag{2.81}$$

However, to extend this to a map

$$\psi_*^{(k,l)} : T^{(k,l)}M \to T^{(k,l)}N, \tag{2.82}$$

we need ψ to be *invertible* (with smooth inverse), in which case we may as well take $N = M$ and assume that $\psi : M \to M$ is a diffeomorphism. In that case, we have

$$(\psi_*^{(k,l)}(\tau_x))(X_1(\psi(x)),\ldots,\theta^l(\psi(x))) := \tau_x(\psi_*^{-1}(X_1(\psi(x))),\ldots,\psi^*(\theta^l(\psi(x)))); \tag{2.83}$$

$$\psi_*^{(k,l)}(\tau_x) = \tau^{j_1\cdots j_l}_{i_1\cdots i_k}(x)\cdot(\psi^{-1})^*_x(dx^{i_1})\otimes\cdots\otimes\psi'_x(\partial_{j_l}). \tag{2.84}$$

This can also be done with ψ replaced by ψ^{-1}, giving maps

$$\psi^*_{(k,l)} : T^{(k,l)}M \to T^{(k,l)}N, \tag{2.85}$$

which in turn induce maps on the sections

$$\psi^*_{(k,l)} : \mathfrak{X}^{(k,l)}(M) \to \mathfrak{X}^{(k,l)}(M), \tag{2.86}$$

often just called ψ^*, via

$$(\psi^*_{(k,l)}\tau)_x(X_1(x),\ldots,\theta^l(x)) = \tau_{\psi(x)}(\psi_*(X_1(x)),\ldots,(\psi^{-1})^*(\theta^l(x))). \tag{2.87}$$

In particular, $\psi^*_{(1,0)}$ is the map ψ^* from (2.58), whereas $\psi^*_{(0,1)} = \psi_*^{-1}$ (recall that $\psi_* \equiv \psi'$).

A natural operation on tensors, which is often used in GR, is ***tensoring***: if

$$\tau_1 \in \mathfrak{X}^{(k_1,l_1)}(M) \qquad \text{and} \qquad \tau_2 \in \mathfrak{X}^{(k_2,l_2)}(M), \tag{2.88}$$

then

$$\tau_1 \otimes \tau_2 \in \mathfrak{X}^{(k_1+k_2,l_1+l_2)}(M) \tag{2.89}$$

is defined by ***concatenation***, i.e.

$$\begin{aligned}&\tau_1\otimes\tau_2(X_1,\dots,X_{k_1},Y_1,\dots Y_{k_2};\theta^1,\dots,\theta^{l_1},\rho^1,\dots,\rho^{l_2}):=\\&\tau_1(X_1,\dots,X_{k_1};\theta^1,\dots,\theta^{l_1})\cdot\tau_2(Y_1,\dots Y_{k_2};\rho^1,\dots,\rho^{l_2}).\end{aligned}\tag{2.90}$$

Indeed, $\mathfrak{X}^{(k,l)}(M)$ itself arose in this way by tensoring copies of $\mathfrak{X}^{(1,0)}(M)$ and $\mathfrak{X}^{(0,1)}(M)$.

Another important operation for GR is ***(index) contraction***: If $k>0$ and $l>0$, then a tensor $\tau\in\mathfrak{X}^{(k,l)}(M)$ may be contracted along one fixed upper and one lower index, say i and j (the result depends on this choice) so as to obtain a tensor $\sigma\in\mathfrak{X}^{(k-1,l-1)}(M)$ with two indices less. Let (e_a) be a basis of T_xM, with ***dual basis*** (ω^a) of T_x^*M (i.e. $\omega^a(e_b)=\delta^a_b$); in local coordinates one could take the (∂_i) basis, with dual (dx^i). Then

$$\sigma^{b_1,\dots,\hat{b}_i,\dots,b_l}_{a_1,\dots,\hat{a}_j,\dots,a_k}(x):=\tau^{b_1,\dots,a,\dots,b_l}_{a_1,\dots,a,\dots,a_k}(x),\tag{2.91}$$

where, according to the Einstein summation convention, a is summed over, and a hat means that the given index is omitted. This is easily seen to be independent of the basis.

Finally, the ***Lie derivative*** $\mathscr{L}_X$, so far only defined on vector fields, may be extended to a linear (and "$C^\infty(M)$-Leibnizian") map

$$\mathscr{L}_X^{(k,l)}:\mathfrak{X}^{(k,l)}(M)\to\mathfrak{X}^{(k,l)}(M)\tag{2.92}$$

in two equivalent ways:

- *Concretely*, writing $\mathscr{L}_X$ for $\mathscr{L}_X^{(k,l)}$ for simplicity, one may define

$$\mathscr{L}_X\tau:=\lim_{t\to0}(\psi_t^*(\tau)-\tau)/t\qquad\qquad(\tau\in\mathfrak{X}^{(k,l)}M),\tag{2.93}$$

  cf. (2.34). In local coordinates, this gives the following explicit formula:

$$\begin{aligned}(\mathscr{L}_X\tau)^{j_1\cdots j_l}_{i_1\cdots i_k}&=X^i\partial_i\tau^{j_1\cdots j_l}_{i_1\cdots i_k}+(\partial_{i_1}X^i)\tau^{j_1\cdots j_l}_{i\cdots i_k}+\cdots+(\partial_{i_n}X^i)\tau^{j_1\cdots j_l}_{i_1\cdots i}\\&\quad-(\partial_jX^{j_1})\tau^{j\cdots j_l}_{i_1\cdots i_k}-\cdots-(\partial_jX^{j_l})\tau^{j_1\cdots j}_{i_1\cdots i_k},\end{aligned}\tag{2.94}$$

  of which (2.8) is clearly a special case.

- *Axiomatically*, one may define the $\mathscr{L}_X$ as the unique linear maps satisfying the rules:

  1. $\mathscr{L}_X^{(0,0)}f=Xf$ for functions $f\in C^\infty(M)\equiv\mathfrak{X}^{(0,0)}M$;
  2. $\mathscr{L}_X^{(0,1)}Y=[X,Y]$ for vector fields $Y\in\mathfrak{X}(M)\equiv\mathfrak{X}^{(0,1)}M$;
  3. $(\mathscr{L}_X^{(1,0)}\theta)(Y)=\mathscr{L}_X(\theta(Y))-\theta(\mathscr{L}_XY)$ for covector fields $\theta\in\Omega(M)\equiv\mathfrak{X}^{(1,0)}M$;
  4. $\mathscr{L}_X^{(k,l)}(\sigma\otimes\tau)=(\mathscr{L}_X\sigma)\otimes\tau+\sigma\otimes\mathscr{L}_X\tau$ for all higher-order tensors (***Leibniz rule***).

It follows from either (a)–(d) or (2.94) that for all cases $\mathscr{L}_X^{(k,l)}\equiv\mathscr{L}_X$ one has the lovely rule

$$[\mathscr{L}_X,\mathscr{L}_Y]=\mathscr{L}_{[X,Y]}.\tag{2.95}$$

2.6 Manifolds with boundaries and corners

For the action principle in GR as well as for things like Penrose diagrams or Cauchy horizons we will need an extension of the manifold concept defined in §2.1 so as to incorporate (smooth) boundaries and, sometimes, corners (which lead to non-smooth boundaries).[135] For $n \geq 1$, let

$$\mathbb{R}^n_+ := \{x \in \mathbb{R}^n \mid x^n \geq 0\}; \qquad \tilde{\mathbb{R}}^n_+ := \{x \in \mathbb{R}^n \mid x^1 \geq 0, \ldots, x^n \geq 0\}. \tag{2.96}$$

Definition 2.8 1. *A C^k-**manifold with boundary** M is defined in the same way as a manifold (cf. §2.1), except that one replaces $\mathbb{R}^n$ by $\mathbb{R}^n_+$ throughout the definition. In particular, each point $x \in M$ has a nbhd $U \in \mathscr{O}(M)$ that is homeomorphic to some $V \in \mathscr{O}(\mathbb{R}^n_+)$.*

2. *Similarly, a **manifold with corners** is defined from the model space $\tilde{\mathbb{R}}^n_+$ instead of $\mathbb{R}^n$.*

3. *In these definitions, C^k-regularity of the transition functions $\varphi_\beta \circ \varphi_\alpha^{-1}$ (see Definition 2.1.4 in §2.1) is defined by declaring $F : V \to \mathbb{R}^m$, where $V \in \mathscr{O}(\mathbb{R}^n_+)$ or $V \in \mathscr{O}(\tilde{\mathbb{R}}^n_+)$, to be C^k, $0 \leq k \leq \infty$, iff F can be extended to a C^k map on some open nbhd of V in $\mathbb{R}^n$.*[136]

4. *In both cases a map $f : M \to \mathbb{R}$ is C^k iff the map $f \circ \varphi_\alpha^{-1} : V_\alpha \to \mathbb{R}$ is C^k for each α.*

5. *The **boundary** ∂M of a manifold M with boundary or corners is the set of all $x \in M$ whose image $\varphi(x)$ in some chart (U, φ) with $x \in U$ lies on the (topological) boundary $\partial\varphi(U)$ of $\varphi(U)$ in $\mathbb{R}^n$ (this is independent of the chart).*[137] *In addition, a boundary point of a manifold with corners is a **corner point** if at least two of the coordinates of $\varphi(x)$ vanish.*

6. *The **interior** $\mathrm{int}(M)$ is defined as $M \backslash \partial M$.*

7. *For $k = \infty$, the **tangent bundle** is defined exactly as in Definition 2.4. In particular, for any $x \in M$, the tangent space T_xM is the space of all point derivations (2.2) of $C^\infty(M)$.*

The boundary of a manifold *with boundary* is itself a manifold (without boundary or corners), in the same class C^k as M itself, of dimension $n-1$ (i.e. one less than M). This should be clear for $\mathbb{R}^n_+$ itself, where $\partial\mathbb{R}^n_+ = \{x \in \mathbb{R}^n \mid x^n = 0\}$, which is clearly $\cong \mathbb{R}^{n-1}$. However, corner points typically ruin C^k regularity of the boundary; removing them leaves a disconnected C^k boundary. On the other hand, in both cases $\mathrm{int}(M)$ is again a "plain", n-dimensional manifold.[138]

Surprisingly, for $M = \mathbb{R}^n_+$ the tangent space is just $T_x\mathbb{R}^n_+ = T_x\mathbb{R}^n$ even at $x \in \partial\mathbb{R}^n_+$, and also for general M the fibers T_xM are vector spaces with a coordinate basis $(\partial/\partial x^1, \ldots, \partial/\partial x^n)$ at any $x \in M$. This makes it possible to define tensors and (semi) Riemannian metrics as usual.

To recover the intuition that tangent vectors at boundary points $x \in \partial M$ should be directed inwards (at least without corners), note that the set-theoretic complement $T_xM \backslash T_x\partial M$ of $T_x\partial M$ is the disjoint union of two open half-spaces of which one, call it T^i_xM, consisting of inward tangent vectors, is distinguished by the property that for any $X \in T^i_xM$ there exists a smooth (or C^k) curve $c : [0, \varepsilon) \to M$ for which $c(0, \varepsilon) \in \mathrm{int}(M)$ and $Xf(x) = df(c(t))/dt_{|t=0}$, as usual.

[135] See Lee (2012) for both boundaries and corners, and Gallot, Hulin, & Lafontaine (1990) for boundaries. Manifolds with corners are usually studied using the *b*-calculus of Melrose (1996).

[136] For $k = \infty$, ***Seeley's extension theorem*** states that this is equivalent to all derivatives of F being bounded on all bounded subsets of the (topological) interior $\mathrm{int}(V)$ of V (Seeley, 1964). See also Grieser (2000).

[137] Either $x \in M$ has an open nbhd $U \cong V \in \mathscr{O}(\mathbb{R}^n)$, in which case $x \notin \partial M$, or it doesn't, in which case $x \in \partial M$.

[138] A basic result is the ***collar neighbourhood theorem***, which states that if M is a smooth manifold with boundary, then ∂M has an open nbhd in M that is diffeomorphic to $\partial M \times [0, 1)$. See e.g. Schultz (undated).

2.7 Summary

- Differential geometry gets going as soon as we define the space $C^\infty(M)$ of smooth (real-valued) functions on a manifold M; this is done through local charts $\varphi : U \to \mathbb{R}^n$ ($U \subset M$).
- The ***coordinates*** $(x^1,\dots,x^n)$ of $x \in U$ with respect to $\varphi = (\varphi^1,\dots,\varphi^n)$ are $x^i = \varphi^i(x)$.
- The ***tangent bundle*** TM to M is the union $TM = \sqcup_{x\in M} T_xM$, where T_xM is the space of ***point derivations*** at x, defined as linear maps $\delta_x : C^\infty(M) \to \mathbb{R}$ that satisfy the Leibniz rule $\delta_x(fg) = \delta_x(f)g(x) + f(x)\delta_x(g)$. Each δ_x takes the form $\delta_x(f) = \frac{d}{dt} f(\gamma(t))_{|t=0}$, where $\gamma : (-\varepsilon,\varepsilon) \to M$ is some curve through $x = \gamma(0)$; then δ_x is called a ***tangent vector*** X_x.
- A smooth section $x \mapsto \delta_x$ of TM corresponds to a ***derivation*** $\delta : C^\infty(M) \to C^\infty(M)$, i.e. a linear map satisfying $\delta(fg) = \delta(f)g + f\delta(g)$. Conversely, each derivation defines point derivations $\delta_x(f) = \delta(f)(x)$. Seen as $x \mapsto X_x$, a derivation $\delta \equiv X$ is a ***vector field*** on M. The set of all vector fields on M is denoted by $\mathfrak{X}(M)$. It is naturally a $C^\infty(M)$ module.
- The coordinates of $X_x \in T_xM$ with respect to φ are $X^i = X_x\varphi^i$ (where $X_x = \delta_x$ is restricted to U), and one has $X_x = X^i\partial_i$, where $\partial_i = \partial/\partial x^i$, and $(\partial_1,\dots,\partial_n)$ form a basis of T_xM.
- The ***cotangent bundle*** T^*M to M is the union $T^*M = \sqcup_{x\in M} T_x^*M$, where T_x^*M is the linear dual $\mathrm{Hom}(T_xM,\mathbb{R})$. Each $C^\infty(M)$-linear map $\theta : \mathfrak{X}(M) \to C^\infty(M)$, called a ***1-form***, comes from a cross-section $x \mapsto \theta_x$ with $\theta_x \in T_x^*M$. The set of all 1-forms on M is called $\Omega(M)$.
- The ***exterior derivative*** $d : C^\infty(M) \to \Omega(M)$ is canonically defined by $df(X) := X(f)$.
- The coordinates of $\theta_x \in T_x^*M$ with respect to φ are $\theta_i = \theta(\partial_i)$, and then $\theta = \theta_i dx^i$.
- The vector bundle $T^{(k,l)}M = \cup_x T_x^{(k,l)}M$ of ***tensors of type*** (k,l) over M is defined by

$$T_x^{(k,l)}M := \mathrm{Hom}((T_xM)^k \times (T_x^*M)^l, \mathbb{R}) \cong (\otimes^k T_x^*M) \otimes (\otimes^l T_xM).$$

  The cross-sections $x \mapsto \tau_x \in T_x^{(k,l)}M$) are the maps $\tau : \mathfrak{X}(M)^k \times \Omega(M)^l \to C^\infty(M)$ that are $k+l$-multilinear linear over $C^\infty(M)$. These maps, also called tensors, form $\mathfrak{X}^{(k,l)}(M)$.
- Important special cases are: $T^{(1,0)} = T^*M$, so that $\mathfrak{X}^{(1,0)}(M) = \Omega(M)$, and $T^{(0,1)} = TM$, so that $\mathfrak{X}^{(0,1)}(M) = \mathfrak{X}(M)$. Furthermore, the metric tensor g of GR will be in $\mathfrak{X}^{(2,0)}(M)$.
- The coordinates $\tau_{i_1\cdots i_k}^{j_1\cdots j_l}$ of $\tau_x \in T_x^{(k,l)}M$ are given by $\tau_x(\partial_{i_1},\dots,\partial_{i_k};dx^{j_1}\dots,dx^{j_l})$, and we have $\tau_x = \tau_{i_1\cdots i_k}^{j_1\cdots j_l}(x)\, dx^{i_1}\otimes\cdots dx^{i_k}\otimes\partial_{j_1}\otimes\cdots\otimes\partial_{j_l}$. For the metric, this gives $g_{ij} = g(\partial_i,\partial_j)$.
- For each vector field $X \in \mathfrak{X}(M)$, the ***Lie derivative*** $\mathscr{L}_X : \mathfrak{X}^{(k,l)}(M) \to \mathfrak{X}^{(k,l)}(M)$ is a linear map that satisfies $\mathscr{L}_X(f\tau) = (Xf)\tau + f\mathscr{L}_X(\tau)$ for each $f \in C^\infty(M)$ and $\tau \in \mathfrak{X}^{(k,l)}(M)$.
- The Lie derivative satisfies three important properties: for vector fields $Y \in \mathfrak{X}(M)$ one has $[\mathscr{L}_X,\mathscr{L}_Y] = \mathscr{L}_{[X,Y]}$ as well as $\mathscr{L}_XY = [X,Y]$, whilst $\mathscr{L}_Xf = Xf$ on functions $f \in C^\infty(M)$.
- Unless stated otherwise, all maps between smooth objects are required to be smooth.
- The ***Einstein summation convention*** holds: *repeated (diagonal) indices are summed over.*

3 Metric differential geometry

The main object of study in GR is the ***metric tensor*** $g \in \mathfrak{X}^{(2,0)}(M)$. This is a smooth family

$$g_x : T_xM \times T_xM \to \mathbb{R} \tag{3.1}$$

of metrics as defined in §2.3, where now $V = T_xM$ is indexed by $x \in M$. Thus, repeating the definition, each g_x is *bilinear*, *symmetric* (i.e. $g_x(X_x,Y_x) = g_x(Y_x,X_x)$ for all $X_x,Y_x \in T_xM$), and *nondegenerate* (i.e. $g_x(X_x,Y_x) = 0$ for all $Y_x \in T_xM$ iff $X_x = 0$). *It need not be positive definite*.

The orthonormal basis $(e_a(x)) = (e_1(x),\dots,e_n(x))$ in which g_x is diagonal (cf. §2.3) may–and typically will–depend on x. But if M is connected, the signature of g_x is independent of x by continuity. Even if M is not connected, we assume it is independent of x. Thus the signature is an intrinsic property not only of each pointwise metric g_x, but even of an entire metric tensor g.

A manifold M with a metric tensor g is called ***semi-Riemannian***, with two special cases:

1. The metric (or manifold) is called ***Riemannian*** if the signature is $(+\cdots+)$. Thus each g_x is positive definite. Given the assumption of symmetry, this *implies* that g_x is nondegenerate, so a metric tensor is Riemannian iff each g_x is symmetric and positive definite.[139]

2. The metric (or manifold) is called ***Lorentzian*** if $\dim(M) = 4$ and $n_- = 1$, i.e. the signature of g is $(-+++)$.[140] Hence with respect to an orthonormal basis (e_a) we have

$$g(e_a,e_b) = \eta_{ab}; \qquad \eta := \mathrm{diag}(-1,1,1,1). \tag{3.2}$$

With some abuse of notation, the symbol η_n, with $\eta \equiv \eta_4$, is also used for the ***Minkowski metric***

$$\eta_n : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}; \qquad \eta_n(X,Y) := \eta_{ab}X^aY^b = -X^0Y^0 + \textstyle\sum_{i=1}^{n-1} X^iY^i, \tag{3.3}$$

where (X^μ) and (Y^μ) are either meant to be Cartesian coordinates on $\mathbb{R}^4$ seen as a vector space,[141] or, identifying $T_x\mathbb{R}^4 \cong \mathbb{R}^4$, denote components of tangent vectors $X = X^\mu\partial_\mu$ etc. with respect to the basis $(\partial_0,\partial_1,\partial_2,\partial_3)$ defined by the usual coordinates (x) on $\mathbb{R}^4$, seen as our manifold M. Either way, $(\mathbb{R}^4,\eta)$, often written as $(\mathbb{M},\eta)$, is ***Minkowski space-time***, which is the oldest and simplest example of a Lorentzian manifold. The fact that, in this special case, the metric is defined not only on each tangent space T_xM, as always, but also on M itself, has no analogue for general Lorentzian manifolds. In special relativity, however, lightcones and other causal structures are defined in $M = \mathbb{R}^4 = \mathbb{M}$ itself, which makes it useful to define the metric η on both M and T_xM. Causal theory for general Lorentzian manifolds will be developed in §5.3.

Lorentzian manifolds underlie GR, but we often invoke examples from Riemannian geometry in order to explain some contrast with the Lorentzian case. Furthermore, Riemannian submanifolds of M are often important, e.g. in the Cauchy problem for GR (see chapter 7).

---

[139] The case $(-\cdots-)$ may also be included here, since an overall change of sign in g makes it Riemannian.

[140] This name is sometimes also used in any dimension $d \geq 2$ provided $n_- = \dim(M) - 1$. Furthermore, a similar comment as in the previous footnote applies: we may as well take $n_+ = \dim(M) - 1$. In any $d \geq 2$, a necessary and sufficient condition for a metrizable manifold M to support a Lorentzian metric is that M is either non-compact, or, if it is compact, has zero Euler characteristic. These conditions are equivalent to the existence of a non-vanishing continuous vector field on M (Palomo & Romero 2006, §1.1; Minguzzi, 2019, §1.8). For deeper topological constraints imposed by Lorentzian metrics with additional (causality) properties, see Chernov & Nemirovski (2013). But in GR one often starts with a metric defined by some formula and looks for a manifold supporting it!

[141] Seen as Minkowski space-time, it is conventional to relabel the usual coordinates of $\mathbb{R}^4$ as (x^0,x^1,x^2,x^3), where $x^0 = t$ denotes time. In diagrams, the time axis is always drawn vertically. We also introduce a convention often used in the (physics) literature: Greek indices μ,ν etc. run from 0 to 3, whereas Latin indices i,j etc. run from 1 to 3. Both Greek and Latin indices midway in the alphabet usually refer to the canonical coordinate basis $\partial_\mu = \partial/\partial x^\mu$ or $\partial_i = \partial/\partial x^i$, whereas indices a,b etc. typically refer to other bases (e_a), often orthonormal ones.

3.1 Lowering and raising indices

Let (M,g) be a (semi) Riemannian manifold. Since each g_x is a metric, the distinction between vectors and covectors is blurred, because as in §2.3 we have "musical" isomorphisms

$$\flat_x : T_xM \to T_x^*M; \qquad \flat_x(X) \equiv X^\flat; \qquad X^\flat(Y) := g_x(X,Y); \tag{3.4}$$

$$\sharp_x : T_x^*M \to T_xM; \qquad \sharp_x(\theta) \equiv \theta_\sharp; \qquad g_x(\theta_\sharp, X) := \theta(X), \tag{3.5}$$

which are each other's inverse. These pointwise isomorphisms induce mutually inverse maps

$$\flat : \mathfrak{X}(M) \to \Omega(M); \qquad \sharp : \Omega(M) \to \mathfrak{X}(M), \tag{3.6}$$

by pointwise application. This leads to the ***lowering and raising of indices***, which is crucial to almost any computation in GR. At any point x (which we omit) we define (g^{ab}) as the inverse (matrix) to (g_{ab}), where $g_{ab} = g(e_a, e_b)$ in some basis e_a (so that $g^{ab}g_{bc} = \delta^a_c$). Then

$$X^\flat_a = g_{ab}X^b; \qquad \theta^a_\sharp = g^{ab}\theta_b, \tag{3.7}$$

which notation may then be extended to any tensor, *where the "sharp" and "flat" signs are usually omitted*. For example, (3.7) is simply written as $X_a = g_{ab}X^b$ and $\theta^a = g^{ab}\theta_b$.

The above definition of (g^{ab}) is consistent with the following one. Extending $\sharp_x$ to a map

$$\sharp_x \otimes \sharp_x : T_x^*M \otimes T_x^*M \to T_xM \otimes T_xM \tag{3.8}$$

in the obvious way, i.e., by linear extension of $\theta \otimes \eta \mapsto \theta_\sharp \otimes \eta_\sharp$, we obtain

$$\sharp_x \otimes \sharp_x(g_x) \in T_x^{(0,2)}M = \mathrm{Hom}(T_x^*M \times T_x^*M, \mathbb{R}). \tag{3.9}$$

If (ω^a) is the dual basis to (e_a), then $g_x^{ab} = \sharp_x \otimes \sharp_x(g_x)(\omega^a(x), \omega^b(x))$, as the reader will verify.

More generally, lowering and raising of specified indices are maps defined, respectively, by

$$\flat : \mathfrak{X}^{(k,l)}(M) \to \mathfrak{X}^{(k+1,l-1)}(M); \qquad \sharp : \mathfrak{X}^{(k,l)}(M) \to \mathfrak{X}^{(k-1,l+1)}(M), \tag{3.10}$$

provided $l > 0$ in the first and $k > 0$ in the second case. Taking the first index for example gives

$$T^\flat(X_1,\ldots,X_{k+1};\theta^1,\ldots,\theta^{l-1}) = T(X_2,\ldots,X_{k+1};X_1^\flat,\theta^1,\ldots,\theta^{l-1}); \tag{3.11}$$

$$T_\sharp(X_1,\ldots,X_{k-1};\theta^1,\ldots,\theta^{l+1}) = T(\theta^1_\sharp,X_1,\ldots,X_{k-1};\theta^2,\ldots,\theta^{l+1}). \tag{3.12}$$

Curvature will described by the Riemann tensor $R \in \mathfrak{X}^{(3,1)}(M)$, of which the only upper index is usually written first. This index may then be lowered, so that $R^\flat \in \mathfrak{X}^{(4,0)}(M)$ has components

$$R^\flat_{abcd} \equiv R_{abcd} = g_{ae}R^e_{bcd}. \tag{3.13}$$

The contraction process explained at the end of the previous chapter, which in principle has nothing to do with the metric, may now elegantly be rewritten in terms of the metric by, e.g.,

$$R_{ab} = R^c_{acb} = g^{cd}R^\flat_{dacb} \equiv g^{cd}R_{dacb}. \tag{3.14}$$

Metric contraction may be done even in case where the original version does not apply, as in

$$R = R^a_{\sharp a} = g^{ab}R_{ab}. \tag{3.15}$$

If $R \in \mathfrak{X}^{(3,1)}(M)$ is the Riemann tensor, so that its first contraction $R \in \mathfrak{X}^{(2,0)}(M)$ is the Ricci tensor, this second contraction yields the ***Ricci scalar***, which again plays a central role in GR.[142]

[142] Our use of the same letter R for the Riemann tensor, the Ricci tensor, and the Ricci scalar will never lead to confusion, as all relevant instances contain indices distinguishing them. For experts: we do *not* use Penrose's abstract index notation, which may clarify things but ever so often leads to typographically awkward expressions.

3.2 Geodesics

Intuitively, geodesics are paths of shortest lengths between two given points.[143] This idea only makes direct sense in the Riemannian case (as opposed to the semi-Riemannian case), with which we therefore start. We will then find a redefinition of a geodesic that does make sense also on semi-Riemannian manifolds. *Throughout this section* (M,g) is a *Riemannian* manifold. It will now be convenient to use *closed* intervals $I=[a,b]$ as the domains of curves $\gamma: I\to M$.

1. The ***length*** of a curve $\gamma:[a,b]\to M$ is defined as

$$L(\gamma) := \int_a^b dt\, \sqrt{g_{\gamma(t)}(\dot\gamma(t),\dot\gamma(t))} \equiv \int_a^b dt\, \|\dot\gamma(t)\|, \tag{3.16}$$

   where $\dot\gamma(t)\in T_{\gamma(t)}M$ is the tangent vector to the curve, cf. (2.25). So in coordinates one has $\gamma(t)=(\gamma^1(t),\dots,\gamma^n(t))$, where $\gamma^i:[a,b]\to\mathbb{R}$, and hence

$$g_{\gamma(t)}(\dot\gamma(t),\dot\gamma(t)) = g_{ij}(\gamma(t))\frac{d\gamma^i(t)}{dt}\frac{d\gamma^j(t)}{dt} \equiv g_{ij}(\gamma(t))\dot\gamma^i(t)\dot\gamma^j(t). \tag{3.17}$$

   Using a change of variables in the integral (3.16), it is easy to show that the length of γ is independent of its parametrization, so that it only depends on the image $\gamma([a,b])$ in M.

2. If M is connected, any two points can be connected by a smooth curve, and hence we can define the ***distance*** $d(x,y)$ between $x,y\in M$ as the infimum of $L(\gamma)$ over all smooth curves $\gamma:[0,1]\to M$ with $\gamma(0)=x$ and $\gamma(1)=y$ (one may equivalently use piecewise smooth curves here, since these can be smoothened, cf. Lemma 5.8 below). This is a metric on M, whose metric topology coincides with the original topology of M.[144]

3. A ***geodesic*** is a curve of extremal length (with a specific parametrization, see below).

We will not precisely explain what this problem in the calculus of variations means, since our goal is merely to motivate Definition 3.1 below, which also applies to the semi-Riemannian case. Therefore, we just outline how this extremal problem is solved. In general, a functional

$$S(\gamma) = \int_a^b dt\, \mathscr{L}(\gamma(t),\dot\gamma(t)) \tag{3.18}$$

is minimized or maximized by some curve γ iff the ***Euler–Lagrange equations*** hold:

$$\frac{d}{dt}\frac{\partial\mathscr{L}}{\partial\dot\gamma^i} - \frac{\partial\mathscr{L}}{\partial\gamma^i} = 0. \tag{3.19}$$

Short of giving an introduction to the calculus of variations, here is a heuristic derivation of (3.19). Let $\gamma_s(t)$ a family of curves indexed by s, such that endpoints are fixed, that is,

$$\gamma_s(a)=\gamma(a); \qquad\qquad \gamma_s(b)=\gamma(b). \tag{3.20}$$

[143] Recall our standing assumption that all maps, including curves and metrics, are smooth. Uniqueness and variational properties of geodesics change completely if the metric is just C^1 (Hartman & Wintner, 1951; Hartman, 1983). On the other hand, most of the smooth theory is already valid in the Hölder class $C^{2,1}$ (Minguzzi, 2015a).

[144] See Jost (2002), pp. 14–15. We do not prove this since it is practically irrelevant for the Lorentzian case.

The extremality condition that defines the variational problem is

$$dS(\gamma_s)/ds = 0. \tag{3.21}$$

On repeatedly using the chain rule and a partial integration, eq. (3.21) with (3.20) gives

$$\begin{aligned}\frac{dS(\gamma_s)}{ds} &= \int_a^b dt \left(\frac{\partial \mathscr{L}}{\partial \gamma_s^i}\frac{\partial \gamma_s^i}{\partial s} + \frac{\partial \mathscr{L}}{\partial \dot\gamma_s^i}\frac{\partial \dot\gamma_s^i}{\partial s}\right) = \int_a^b dt \left(\frac{\partial \mathscr{L}}{\partial \gamma_s^i}\frac{\partial \gamma_s^i}{\partial s} + \frac{\partial \mathscr{L}}{\partial \dot\gamma_s^i}\frac{\partial}{\partial t}\frac{\partial \gamma_s^i}{\partial s}\right) \\ &= \int_a^b dt \left(\frac{\partial \mathscr{L}}{\partial \gamma_s^i} - \frac{\partial}{\partial t}\frac{\partial \mathscr{L}}{\partial \dot\gamma_s^i}\right)\frac{\partial \gamma_s^i}{\partial s} + \left|_b^a \frac{\partial \mathscr{L}}{\partial \dot\gamma_s^i}\frac{d\gamma_s^i}{ds}\right.. \end{aligned} \tag{3.22}$$

Then (3.20) gives $d\gamma_s(a)/ds = d\gamma_s(b)/ds = 0$, so that, for arbitrary γ_s and hence arbitrary $\partial\gamma_s/\partial s$, eq. (3.21) implies (3.19), in which s is dropped and hence $\partial/\partial t$ becomes d/dt.

The Euler–Lagrange equations for the length functional (3.16) are not very nice, but they can be simplified if a preferred ("affine") parametrization is used. To motivate this, instead of the length (3.16), we now start from the (kinetic) ***energy*** of our curve γ, defined as

$$E(\gamma) := \int_a^b dt\, g_{\gamma(t)}(\dot\gamma(t),\dot\gamma(t)) = \int_a^b dt\, \|\dot\gamma(t)\|^2. \tag{3.23}$$

For the energy (3.23), the Euler–Lagrange equations (3.19) give the ***geodesic equation***

$$\ddot\gamma^i(t) + \Gamma^i_{jk}(\gamma(t))\dot\gamma^j(t)\dot\gamma^k(t) = 0, \tag{3.24}$$

or briefly $\ddot\gamma^i + \Gamma^i_{jk}\dot\gamma^j\dot\gamma^k = 0$, where $\ddot\gamma = d^2\gamma/dt^2$, and the ***Christoffel symbols*** are given by

$$\Gamma^i_{jk} := \tfrac{1}{2}g^{im}(g_{mj,k} + g_{mk,j} - g_{jk,m}), \tag{3.25}$$

where we have introduced another useful notational convention from GR:

$$\tau^{j_1\cdots j_l}_{i_1\cdots i_k,j} = \partial_j \tau^{j_1\cdots j_l}_{i_1\cdots i_k}. \tag{3.26}$$

Warning: the Christoffel symbols do *not* form the components of a would-be tensor "$\Gamma \in \mathfrak{X}^{(2,1)}(M)$": physicists see this from their incorrect behaviour under coordinate transformations, whereas mathematicians note that Γ fails the tensoriality test, cf. Proposition 2.7. We will see, however, that the Γ-*symbols* do combine into the Riemann *tensor*!

To derive (3.24) for (3.23), i.e., for $\mathscr{L}(\gamma(t),\dot\gamma(t)) = g_{ij}(\gamma(t))\dot\gamma^i(t)\dot\gamma^j(t)$, one uses

$$\frac{\partial \mathscr{L}}{\partial \gamma^i} = g_{jk,i}\dot\gamma^j\dot\gamma^k; \tag{3.27}$$

$$\frac{d}{dt}\frac{\partial \mathscr{L}}{\partial \dot\gamma^i} = 2\frac{d}{dt}g_{ij}\dot\gamma^j = 2(g_{ij,k}\dot\gamma^k\dot\gamma^j + g_{ij}\ddot\gamma^j) = (g_{ij,k}+g_{ik,j})\dot\gamma^k\dot\gamma^j + 2g_{ij}\ddot\gamma^j. \tag{3.28}$$

Whereas solutions of (3.24) extremize the *energy* for any parametrization, for the *length* (3.16), the Euler–Lagrange equations only take the form (3.24) iff $\|\dot\gamma(t)\|$ is constant, in which case the parametrization of the curve $\gamma\colon [a,b] \to M$ is said to be ***affine***. In particular, if $\|\dot\gamma(t)\| = 1$ for all $t \in I$, then we say that γ is parametrized by ***arc length***.

Definition 3.1 *A geodesic is a curve $\gamma : I \to M$ (with $I \subset \mathbb{R}$ connected) that satisfies* (3.24).

On this definition, geodesics still extremize length, but eq. (3.24) *implies* that $\|\dot\gamma(t)\|$ is constant, as can be shown by computing $d(\|\dot\gamma(t)\|^2)/dt$ from (3.17). This time-derivative equals

$$\frac{d\|\dot\gamma(t)\|^2}{dt} = g_{ij,k}\dot\gamma^i\dot\gamma^j\dot\gamma^k + 2g_{ij}\ddot\gamma^i\dot\gamma^j. \tag{3.29}$$

Eliminating $\ddot\gamma^i$ via (3.24) then leads to a cancellation making the right-hand side zero; a neater calculation will be given after (3.66). *The definition of a geodesic therefore depends on the parametrization of* γ: a reparametrized geodesic may no longer satisfy (3.24), except when the reparametrization is *affine*, i.e. $s = at + b$. However, one has the following useful criterion.

Proposition 3.2 *Some curve $\gamma : [a,b] \to M$ can be reparametrized so as to* become *a geodesic iff the right-hand side of* (3.24) *equals $f \cdot \dot\gamma^i$ for some function $f(t)$ defined along γ.*

Proof. If some curve $t \mapsto \gamma(t)$ satisfies (3.24), then $t \mapsto \gamma(s(t))$ satisfies (3.24) with right-hand side $\ddot s\dot\gamma^i$, and conversely one can solve $f(t) = \ddot s(t)$ for s and switch to $\gamma \circ s^{-1}$. □

Such a (poorly parametrized) curve that is "almost" a geodesic is sometimes called a ***pregeodesic***. In $M = \mathbb{R}^n$ with flat metric (i.e. $g_{ij} = \delta_{ij}$) geodesics are straight lines that form *shortest* paths between two given points. This is also true in e.g. hyperbolic space, and it is always true for sufficiently short geodesics. On the sphere (where geodesics are great circles) one has two geodesics between two generic points; but only one has minimal length. These lengths coincide iff the two points are polar opposites, in which case one has infinitely many geodesics. See §5.5.

In the intuitive idea of geodesics the focus is on endpoints, whereas in defining geodesics as solutions to the ODE (3.24) the focus is on the initial point $\gamma(0)$ and initial velocity $\dot\gamma(0)$. The solution to (3.24) is uniquely defined by these data, except for I. But like any solution to an ODE, γ has some maximal domain of definition $I \subset \mathbb{R}$, and this domain may or may not equal $\mathbb{R}$.

Definition 3.3 *If all geodesics $\gamma : I \to \mathbb{R}$ with given initial point $\gamma(0)$ and initial velocity $\dot\gamma(0)$ can be defined on the maximum interval $I = \mathbb{R}$, we say that (M,g) is* **geodesically complete**.

For example, $\mathbb{R}^n$, the sphere S^n, and hyperbolic space H^n are geodesically complete (cf. §4.4). In the Riemannian case this is equivalent to a purely topological property. For $x, y \in M$ define

$$d(x,y) := \inf\{L(\gamma) \mid \gamma : [a,b] \to M, \gamma(a) = x, \gamma(b) = y\}. \tag{3.30}$$

It is easy to show that this defines a metric in the topological sense, i.e. a symmetric function $d : M \times M \to [0,\infty)$ that satisfies $d(x,y) = 0$ iff $x = y$ and $d(x,y) \le d(x,z) + d(z,y)$. In other words, a Riemannian manifold (M,g) is also a metric space (M,d). For the latter, one has the usual notion of completeness in the sense that any Cauchy sequence converges.

Theorem 3.4 **(Hopf-Rinow)** *A Riemannian manifold (M,g) is geodesically complete iff the corresponding metric space (M,d) defined by* (3.30) *is complete. In that case, any two points x, y can be joined by a geodesic of minimum length (compared with all curves from x to y).*

Since this theorem has no analogue in the Lorentzian case we will not prove it. We do note that any compact Riemannian manifold is complete. On the other hand, examples of incomplete Riemannian manifolds are provided by open bounded sets $\Omega \subset \mathbb{R}^n$ with flat metric inherited from $\mathbb{R}^n$, or $\mathbb{R}^n$ itself with one or more points or regions omitted. Such examples also show that in the incomplete case the infimum in (3.30) may not be attained. Many Lorentzian manifolds of interest to GR are geodesically *in*complete in a nontrivial (i.e. inextendible) sense; see chapter 6.

3.3 Linear connections

The definition of a geodesic as a curve γ whose tangent vector $\dot\gamma$ satisfies (3.24) along the curve for which $\gamma(t)$ is defined was inspired by the Riemannian case, but it will be taken as the definition of a geodesic on a semi-Riemannian manifold, too. In support, we now give a geometric perspective on the Christoffel symbols Γ^i_{jk} and hence on the geodesic equation (3.24).

Definition 3.5 *A* **linear connection** *on M (which is the same thing as a connection on the tangent bundle TM, see below), or, equivalently, a* **covariant derivative** *on $\mathfrak{X}(M)$, is a map*

$$X \mapsto \nabla_X : \mathfrak{X}(M) \to \mathfrak{X}(M), \tag{3.31}$$

where X itself is a vector field on M (i.e. $X \in \mathfrak{X}(M)$), such that:

1. *The map $X \mapsto \nabla_X$ is $\mathbb{R}$-linear as well as $C^\infty(M)$-linear, i.e.*

$$\nabla_{fX} Y = f \nabla_X Y \qquad\qquad \forall f \in C^\infty(M); \tag{3.32}$$

2. *The map $Y \mapsto \nabla_X Y$ is $\mathbb{R}$-linear but not $C^\infty(M)$-linear: it satisfies the* **Leibniz rule**

$$\nabla_X(fY) = (Xf)Y + f\nabla_X Y \qquad\qquad \forall f \in C^\infty(M). \tag{3.33}$$

This definition also makes sense on any open $U \in \mathscr{O}(M)$, and in fact if $x \in U$, then $\nabla_X Y(x)$ only depends on the value of X at x and the restriction of Y to U; this follows from (3.32) - (3.33) and the definition of a manifold. Hence we may compute covariant derivatives locally. Recall that a local frame (e_a) for $\mathfrak{X}(U)$ consists of n maps $e_a : U \to TM$ such that at each $x \in U$ the vectors $e_a(x) \in T_xM$ form a basis of T_xM ($a = 1, \dots, n$). The corresponding dual basis (ω^a) for $\Omega(U)$ then consists of the $\omega^a(x) \in T_x^*M$ that satisfy $\omega^a(e_b) = \delta^a_b$. The given connection ∇ is then completely characterized by its ***connection coefficients*** ω^c_{ab}, defined (at each x) by

$$\nabla_{e_a} e_b = \omega^c_{ab} e_c. \tag{3.34}$$

Indeed, from (2.68) - (2.69) we may write $X = X^a e_a$, where $X^a = \omega^a(X) \in C^\infty(U)$, so

$$\begin{aligned} \nabla_X Y &= \nabla_{X^a e_a}(Y^b e_b) = X^a \nabla_{e_a}(Y^b e_b) \\ &= X^a (e_a(Y^b) \cdot e_b + Y^b \nabla_{e_a} e_b) \\ &= X^a (e_a(Y^c) + Y^b \omega^c_{ab}) e_c. \end{aligned} \tag{3.35}$$

We write $\nabla_X Y^a$ for $(\nabla_X Y)^a$, so that $\nabla_X Y = (\nabla_X Y^a) e_a$. We therefore have

$$\nabla_X Y^a = X(Y^a) + \omega^a_{bc} X^b Y^c, \tag{3.36}$$

where $X(Y^a)$ is the action of the vector field X on the *function* $Y^a \in C^\infty(U)$. In terms of a coordinate basis $(e_\mu = \partial_\mu)$, $(\omega^\nu = dx^\nu)$, writing $\nabla_\mu := \nabla_{\partial_\mu}$ the above relations imply

$$\omega^\rho_{\mu\nu} = dx^\rho(\nabla_\mu \partial_\nu); \tag{3.37}$$

$$\nabla_X Y^\rho = X^\mu (\partial_\mu Y^\rho + \omega^\rho_{\mu\nu} Y^\nu); \tag{3.38}$$

$$\nabla_\mu Y^\rho = \partial_\mu Y^\rho + \omega^\rho_{\mu\nu} Y^\nu. \tag{3.39}$$

Linear connections formalize Levi-Civita's notion of ***parallel transport***. It follows from (3.36) or (3.38) that $\nabla_X Y$ only depends on the values of Y along the flow lines of X, for

$$\nabla_X Y^a(x) = \frac{d}{dt} Y^a(\psi_t(x))_{|t=0} + \omega^a_{bc}(x) X^b(x) Y^c(x), \tag{3.40}$$

where ψ is the flow of X. Conversely, given some curve $\gamma : I \to M$ with tangent vectors $\dot\gamma$ (*defined along γ only!*), the covariant derivative $\nabla_{\dot\gamma} Y$ of Y along γ is well defined for any vector field Y defined near or even on $\gamma(I)$ alone; for in (local) coordinates we have

$$\begin{aligned} \nabla_{\dot\gamma} Y^\rho_{\gamma(t)} &= \dot\gamma^\mu(t)(\partial_\mu Y^\rho_{\gamma(t)} + \omega^\rho_{\mu\nu}(\gamma(t)) Y^\nu_{\gamma(t))} \\ &= \frac{d}{dt} Y^\rho_{\gamma(t)} + \omega^\rho_{\mu\nu}(\gamma(t)) \frac{d\gamma^\mu(t)}{dt} Y^\nu_{\gamma(t)}, \end{aligned} \tag{3.41}$$

where $\gamma^\mu : I \to \mathbb{R}$ are the coordinates of the curve (in some given chart), as before.

Definition 3.6 *A (necessarily unique) vector field $t \mapsto Y_{\gamma(t)} \in T_{\gamma(t)}M$ defined along a given curve γ is the* parallel-transport *of some initial vector $Y \in T_{\gamma(0)}M$ along γ if Y satisfies*

$$\nabla_{\dot\gamma} Y = 0. \tag{3.42}$$

This generalizes the Euclidean practice of freely moving vectors in $\mathbb{R}^n$ from place to place, to arbitrary (semi) Riemannian manifolds. The price one pays is that such motions can only be carried out once a linear connection has been defined. The ***flat connection*** on $\mathbb{R}^n$ (with flat metric $g = \delta$), defined in the standard coordinates by $\omega^\rho_{\mu\nu} = 0$ gives $\nabla_\mu = \partial_\mu$ and hence $Y_{\gamma(t)} = Y_{\gamma(0)} = Y$ for all t. Hence "freely moving vectors" in $\mathbb{R}^n$ is *relative to this flat connection*.

Like the Christoffel symbols, the connection coefficients do not form the components of a tensor (the relation between the two will be clarified shortly). However, various tensors may be defined via the connection. For now, we just define the ***torsion*** $\tau_\nabla \in \mathfrak{X}^{(2,1)}(M)$ of ∇ by

$$\tau_\nabla(X, Y, \theta) := \theta(\nabla_X Y - \nabla_Y X - [X, Y]). \tag{3.43}$$

A simple computation shows that this expression is $C^\infty(M)$-linear in each entry, so Proposition 2.7 shows τ is indeed a tensor of the said kind. In the coordinate basis (∂_μ), we have

$$\tau^\rho_{\mu\nu} = \omega^\rho_{\mu\nu} - \omega^\rho_{\nu\mu}, \tag{3.44}$$

since $[\partial_\mu, \partial_\nu] = 0$. Hence the connection ∇ is ***torsion-free*** iff any of the following hold:

$$\omega^\rho_{\mu\nu} = \omega^\rho_{\nu\mu}; \tag{3.45}$$

$$\nabla_\mu \partial_\nu = \nabla_\nu \partial_\mu; \tag{3.46}$$

$$\nabla_X Y - \nabla_Y X = [X, Y]. \tag{3.47}$$

We are now in a position to restate and generalize Definition 3.1:

Definition 3.7 *Given some linear connection ∇ on M, a* **geodesic** *in M is a curve γ for which*

$$\nabla_{\dot\gamma} \dot\gamma = 0. \tag{3.48}$$

That is, the tangent vector $\dot\gamma$ to γ is parallel transported along γ. As before, this definition requires a specific parametrization of γ, which is unique up to affine transformations of t. One has a similar situation as in the metric case for detecting "wrongly parametrized" geodesics:

Proposition 3.8 *Some curve $\gamma:[a,b]\to M$ can be reparametrized so as to* become *a geodesic iff the right-hand side of* (3.48) *equals $f\dot\gamma$, for some function $f(t)$ defined along γ.*

The proof is analogous to Proposition 3.24. Using (local) coordinates, eq. (3.48) may be brought into a form that is strikingly similar to (3.24). Since according to (3.41) with $Y\rightsquigarrow\dot\gamma$ the expression $\dot\gamma^\mu\partial_\mu\dot\gamma^\rho$ is just $d^2\dot\gamma^\rho/dt^2\equiv\ddot\gamma^\rho$, we obtain

$$\ddot\gamma^\rho+\omega^\rho_{\mu\nu}\dot\gamma^\mu\dot\gamma^\nu=0, \tag{3.49}$$

from which it is obvious that geodesics are insensitive to the torsion (3.44) of the connection. Eq. (3.49) looks like the geodesic equation (3.24), with the difference that in (3.49) the coefficients $\omega^\rho_{\mu\nu}$ are defined by (3.37) in terms of an arbitrary linear connection ∇, whereas those in (3.24) are the Christoffel symbols (3.25) defined by the metric. Their relationship is as follows.

Theorem 3.9 **(Levi-Civita)** *Any (semi) Riemannian manifold (M,g) admits a unique linear connection ∇ (called the* **Levi-Civita connection***) that satisfies the following two properties:*

1. *The torsion τ_∇ associated to ∇ vanishes, i.e. $\nabla_XY-\nabla_YX=[X,Y]$.*
2. *The connection ∇ and the metric g are related by the condition that for all $X,Y,Z\in\mathfrak{X}(M)$,*

$$X(g(Y,Z))=g(\nabla_XY,Z)+g(Y,\nabla_XZ). \tag{3.50}$$

These conditions imply that the connection coefficients of ∇ are the Christoffel symbols (3.25)*:*

$$\omega^\rho_{\mu\nu}=\Gamma^\rho_{\mu\nu}. \tag{3.51}$$

As soon as we have extended ∇ to arbitrary tensors, we will see that (3.50) comes down to

$$\nabla_Xg=0\qquad\qquad\forall X\in\mathfrak{X}(M). \tag{3.52}$$

Also, $X(g(Y,Z))$ will be the same as $\nabla_X(g(Y,Z))$, hence some authors elegantly write (3.50) as

$$\nabla_X\langle Y,Z\rangle=\langle\nabla_XY,Z\rangle+\langle Y,\nabla_XZ\rangle. \tag{3.53}$$

Proof. Using (3.47) and (3.50), one computes

$$X(g(Y,Z))-Z(g(X,Y))+Y(g(Z,X)),$$

and rearranges this to obtain the so-called ***Koszul formula***, partly written in the notation (3.53):

$$\langle\nabla_XY,Z\rangle=\tfrac12(X\langle Y,Z\rangle+Y\langle Z,X\rangle-Z\langle X,Y\rangle-\langle X,[Y,Z]\rangle+\langle[X,Y],Z\rangle+\langle Y,[Z,X]\rangle). \tag{3.54}$$

Since g is nondegenerate this uniquely fixes ∇_XY, and in a coordinate basis this gives (3.51).

To prove existence, one easily checks (3.32) and (3.33) from (3.54). Finally, running the derivation of (3.54) from (3.47) and (3.50) backwards verifies (3.47) and (3.50). □

3.4 General connections on vector bundles

For a more general understanding of the above constructions, as well as for a clean extension of linear connections from vector fields to arbitrary tensors (which one often needs in GR), we briefly discuss connections on arbitrary vector bundles. Similar to Definition 3.10, we put:

Definition 3.10 *A* **connection on a vector bundle** $E \to M$ *is a linear map*

$$X \mapsto \nabla_X : \Gamma(E) \to \Gamma(E), \tag{3.55}$$

where $X \in \mathfrak{X}(M)$*, such that:*

1. *The map* $X \mapsto \nabla_X$ *is* $\mathbb{R}$*-linear as well as* $C^\infty(M)$*-linear in* X*, cf.* (3.32)*;*
2. *The map* $s \mapsto \nabla_X s$ *is* $\mathbb{R}$*-linear but not* $C^\infty(M)$*-linear: it satisfies the* **Leibniz rule**

$$\nabla_X(fs) = (Xf)s + f\nabla_X s \quad (f \in C^\infty(M)). \tag{3.56}$$

A linear connection is then a connection (in the above sense) on the tangent bundle. The general story is almost the same, including the localization of $\nabla_X s(x)$ to the flow lines of X arbitrarily close to x, and hence to any $U \in \mathscr{O}(M)$, $x \in U$. In particular, define a local frame (u_{a}), where $\mathsf{a} = 1,\ldots,k = \dim(E_x)$, i.e. the rank of E, by the properties that (i) $u_{\mathsf{a}} \in \Gamma(U,E)$, i.e., the restriction of $\Gamma(E) \equiv \Gamma(M,E)$ to some $U \in \mathscr{O}(M)$; and (ii) the set $u_{\mathsf{a}}(x)_{a=1,\ldots\dim(E_x)}$ forms a basis of E_x for all $x \in U$. This once again yields ***connection coefficients*** defined by

$$\nabla_\mu u_{\mathsf{b}} = C^{\mathsf{c}}_{\mu\mathsf{b}} u_{\mathsf{c}}. \tag{3.57}$$

The difference with the tangent bundle is that the three indices carried by C are no longer of the same type: b and c label basis vectors in E_x, whereas μ refers to the canonical coordinate base of T_xM (recall that $\nabla_\mu = \nabla_{\partial_\mu}$). Writing $s(x) = s^{\mathsf{a}}(x)u_{\mathsf{a}}(x)$, we now have

$$\nabla_\mu s^{\mathsf{a}} = \partial_\mu s^{\mathsf{a}} + C^{\mathsf{a}}_{\mu\mathsf{b}} s^{\mathsf{b}}, \tag{3.58}$$

cf. (3.39). This is often written as

$$\nabla_\mu s = \partial_\mu s + \omega_\mu s, \tag{3.59}$$

in which s is seen as a vector with components s^{a} relative to the given basis (u_{a}) and hence ω_μ is a matrix with components $C^{\mathsf{a}}_{\mu\mathsf{b}}$, or $s \in \Gamma(E)$ and $\omega_\mu(x) \in \mathrm{Hom}(E_x,E_x)$.[145]

A vector bundle E may be equipped with a ***metric***, i.e. nondegenerate symmetric bilinear form $g_x : E_x \times E_x \to \mathbb{R}$ defined for each $x \in M$, that is smooth in x in the sense that for any $s,t \in \Gamma(E)$ the function $g(s,t) : M \to \mathbb{R}$ defined by $x \mapsto g_x(s(x),t(x))$ is smooth. For example, a (semi) Riemannian metric on M is a metric on $E = TM$ in precisely this sense. A connection ∇ on E is then called ***metric*** if for all $s,t \in \Gamma(E)$ we have

$$X(g(s,t)) = g(\nabla_X s,t) + g(s,\nabla_X t). \tag{3.60}$$

[145] Even more abstractly, connections may be regarded as maps $\nabla : \Gamma(E) \to \Gamma(T^*M \otimes E) \equiv \Omega^1(E)$, i.e. the space of E-valued 1-forms, that satisfy $\nabla(fs) = df \otimes s + f\nabla s$; the connection with the main text is $\nabla_X s = \nabla s(X)$. In that case we may write $\nabla = d + \omega$, where $\omega \in \Omega^1(\mathrm{Hom}(E,E))$, i.e. ω is a 1-form taking values in the vector bundle $\mathrm{Hom}(E,E)$. Even more generally (for those familiar with the de Rham complex $\Omega^\bullet(M)$ and its relative $\Omega^\bullet(E)$), we may define $\nabla : \Omega^p(E) \to \Omega^{p+1}(E)$, where $p = 0,\ldots,k$ with $\Omega^0(E) \equiv \Gamma(E)$, as the unique extension of the above map $\nabla : \Omega^0(E) \to \Omega^1(E)$ that satisfies $\nabla(\alpha \otimes s) = d\alpha \otimes s + (-1)^p \alpha \wedge \nabla s$, where $\alpha \in \Omega^p(M)$ and $s \in \Gamma(E)$.

For example, the Levi-Civita connection on TM is obviously metric in this sense.

Furthermore, take $E = T^*M$, and define ∇^* in coordinates through its components by

$$\nabla^*_\mu \theta_\nu := \partial_\mu \theta_\nu - \Gamma^\rho_{\mu\nu}\theta_\rho, \tag{3.61}$$

where the $\Gamma^\rho_{\mu\nu}$ are the Christoffel symbols defined by some (semi) Riemannian metric on M, cf. (3.25). This turns out to be a connection indeed (check the axioms), whose rationale (notably of the minus sign!) is the Leibniz-type property (or product rule)

$$X(\theta(Y)) = (\nabla^*_X\theta)(Y) + \theta(\nabla_X Y), \tag{3.62}$$

which, omitting the star, may look even more elegant in the form

$$\nabla_X\langle\theta, Y\rangle = \langle\nabla_X\theta, Y\rangle + \langle\theta, \nabla_X Y\rangle, \tag{3.63}$$

where by *fiat* we have declared that on functions (such as $\langle\theta, Y\rangle \equiv \theta(Y)$) the covariant derivative ∇_X is simply X, i.e. $\nabla_X f \equiv Xf$, $f \in C^\infty(M)$. Eq. (3.62) or (3.63) might, of course, have been used to define $\nabla^* \equiv \nabla : \Omega(M) \to \Omega(M)$ in the first place, yielding (3.61). In fact, any linear connection defines a dual connection ∇^* on T^*M by (3.62).

Combining (3.39) and (3.61), we define a covariant derivative $\nabla^{(k,l)} : \mathfrak{X}^{(k,l)} \to \mathfrak{X}^{(k,l)}$ by

$$\begin{aligned}(\nabla^{(k,l)}_\mu\tau)^{\rho_1\cdots\rho_l}_{\nu_1\cdots\nu_k} \equiv \nabla^{(k,l)}_\mu \tau^{\rho_1\cdots\rho_l}_{\nu_1\cdots\nu_k} = \partial_\mu \tau^{\rho_1\cdots\rho_l}_{\nu_1\cdots\nu_k} + \Gamma^{\rho_1}_{\mu\sigma}\tau^{\sigma\cdots\rho_l}_{\nu_1\cdots\nu_k} + \cdots + \Gamma^{\rho_l}_{\mu\sigma}\tau^{\rho_1\cdots\sigma}_{\nu_1\cdots\nu_k} \\ - \Gamma^\sigma_{\mu\nu_1}\tau^{\rho_1\cdots\rho_l}_{\sigma\cdots\nu_k} - \cdots - \Gamma^\sigma_{\mu\nu_k}\tau^{\rho_1\cdots\rho_l}_{\nu_1\cdots\sigma}.\end{aligned} \tag{3.64}$$

Those who do not like coordinate definitions "by formula" may be reassured that $\nabla^{(k,l)}$ is the unique connection on $T^{(k,l)}M$ that, similarly to (3.63), satisfies the Leibniz rule

$$\begin{aligned}\nabla_X(\tau(X_1,\ldots,X_k,\theta^1,\ldots,\theta^l)) = (\nabla^{(k,l)}_X\tau)(X_1,\ldots,X_k,\theta^1,\ldots,\theta^l) \\ + \tau(\nabla_X X_1,\ldots,X_k,\theta^1,\ldots,\theta^l) + \cdots + \tau(X_1,\ldots,X_k,\theta^1,\ldots,\nabla^*_X\theta^l),\end{aligned} \tag{3.65}$$

where the case $k = l = 0$ is taken to mean $\nabla^{(0,0)}_X = X$ on $\mathfrak{X}^{(0,0)}(M) = C^\infty(M)$. Eq. (3.65) recovers $\nabla^{(0,1)} = \nabla$ on $\mathfrak{X}^{(0,1)}(M) = \mathfrak{X}(M)$ as well as $\nabla^{(1,0)} = \nabla^*$ on $\mathfrak{X}^{(1,0)}(M) = \Omega(M)$.

This construction of $\nabla^{(k,l)}$ works for any linear connection ∇. If the latter is the Levi-Civita connection, then (3.65) implies that its defining property (3.50) elegantly reads

$$\nabla^{(2,0)}_X g \equiv \nabla_X g = 0. \tag{3.66}$$

As in (3.66), in general one often writes ∇ for any $\nabla^{(k,l)}$, and physicists write (3.66) as

$$g_{\mu\nu;\sigma} = 0, \tag{3.67}$$

using the ***semi-colon notation***, in which $\tau^{\rho_1\cdots\rho_l}_{\nu_1\cdots\nu_k;\mu}$ means $\nabla_\mu\tau^{\rho_1\cdots\rho_l}_{\nu_1\cdots\nu_k}$, much as $\tau^{\rho_1\cdots\rho_l}_{\nu_1\cdots\nu_k,\mu}$ means $\partial_\mu\tau^{\rho_1\cdots\rho_l}_{\nu_1\cdots\nu_k}$. As an application, let us show once again that $d(\|\dot\gamma(t)\|)/dt = 0$ for geodesics γ:

$$\frac{d\|\dot\gamma(t)\|^2}{dt} = \frac{dg(\dot\gamma,\dot\gamma}{dt} = \dot\gamma(g(\dot\gamma,\dot\gamma)) = (\nabla_{\dot\gamma}g)(\dot\gamma,\dot\gamma) + g(\nabla_{\dot\gamma}\dot\gamma,\dot\gamma) + g(\dot\gamma,\nabla_{\dot\gamma}\dot\gamma),$$

where we used (3.65). Eqs. (3.66) and (3.48) then make the right-hand side $0+0+0=0$.

Alternatively, one may recall the description (2.64) of $T^{(k,l)}M$ as the tensor product of k copies of T^*M and l copies of TM. In general, given two vector bundles $E^{(1)} \to M$ and $E^{(2)} \to M$, with connections $\nabla^{(1)}$ and $\nabla^{(2)}$, there is a unique connection $\nabla^{(1\otimes 2)}$ on the vector bundle tensor product $E^{(1)} \otimes E^{(2)} = \sqcup_{x\in M} E^{(1)}_x \otimes E^{(2)}_x$ that satisfies the product rule

$$\nabla^{(1\otimes 2)}\big(s^{(1)} \otimes s^{(2)}\big) = \nabla^{(1)}\big(s^{(1)}\big) \otimes s^{(2)} + s^{(1)} \otimes \nabla^{(2)}\big(s^{(2)}\big). \tag{3.68}$$

This may be iterated to the tensor product of finitely many vector bundles, and hence (for any linear connection ∇) the connection $\nabla^{(k,l)}$ defined by (3.64) or (3.65) is just the tensor product of the individual connections on each copy of TM or T^*M present in $T^{(k,l)}M$.

It follows from (3.62) that (for any ∇) the connection $\nabla^{(k,l)}$ *commutes with contraction*. Contracting the first upper and lower indices and writing $\sigma^{\rho_2\cdots\rho_l}_{\nu_2\cdots\nu_k} = \tau^{\nu_1\rho_2\cdots\rho_l}_{\nu_1\nu_2\cdots\nu_k}$, one has

$$(\nabla^{(k,l)}_\mu \tau)^{\nu_1\rho_2\cdots\rho_l}_{\nu_1\nu_2\cdots\nu_k} = (\nabla^{(k,l)}_\mu \sigma)^{\rho_2\cdots\rho_l}_{\nu_2\cdots\nu_k}, \tag{3.69}$$

and similarly for any other pair of upper and lower indices. In particular, this makes the physicists' notation $\tau^{\nu_1\rho_2\cdots\rho_l}_{\nu_1\nu_2\cdots\nu_k;\mu}$ unambiguous. For example, for the Ricci tensor (see §4.5) we have

$$R_{\mu\nu;\sigma} = R^\rho_{\mu\rho\nu;\sigma}. \tag{3.70}$$

If ∇ satisfies (3.52), then $\nabla^{(k,l)}$ in addition commutes with contraction in the metric sense explained before (3.15), so that e.g., using (3.67), for the Ricci scalar we have

$$R_{,\sigma} = R_{;\sigma} = (g^{\mu\nu}R_{\mu\nu})_{;\sigma} = g^{\mu\nu}{}_{;\sigma}R_{\mu\nu} + g^{\mu\nu}R_{\mu\nu;\sigma} = g^{\mu\nu}R_{\mu\nu;\sigma}. \tag{3.71}$$

Finally, $\nabla^{(k,l)}$ may be used to rewrite the formula (2.94) for the Lie derivative as

$$\begin{aligned}\mathscr{L}_X \tau^{\rho_1\cdots\rho_l}_{\nu_1\cdots\nu_k} &= \nabla_X \tau^{\rho_1\cdots\rho_l}_{\nu_1\cdots\nu_k} + (\nabla_{\nu_1}X^\nu)\tau^{\rho_1\cdots\rho_l}_{\nu\cdots\nu_k} + \cdots + (\nabla_{\nu_n}X^\nu)\tau^{\rho_1\cdots\rho_l}_{\nu_1\cdots\nu} \\ &\quad - (\nabla_\rho X^{\rho_1})\tau^{\rho\cdots\rho_l}_{\nu_1\cdots\nu_k} - \cdots - (\nabla_\rho X^{\rho_l})\tau^{\rho_1\cdots\rho}_{\nu_1\cdots\nu_k},\end{aligned} \tag{3.72}$$

since all Christoffel symbols cancel out (check!).[146] For example, using (3.52) we obtain

$$\mathscr{L}_X g_{\mu\nu} = (\nabla_\mu X^\rho)g_{\rho\nu} + (\nabla_\nu X^\rho)g_{\mu\rho} = X_{\nu;\mu} + X_{\mu;\nu}. \tag{3.73}$$

A vector field X for which $\mathscr{L}_X g = 0$ is called a ***Killing (vector) field***.[147] Eq. (3.73) gives

$$X_{\nu;\mu} + X_{\mu;\nu} = 0. \tag{3.74}$$

Flows of Killing fields are ***isometries***, that is, diffeomorphisms preserving the metric. In the notation of (2.84), this means that $\psi_t^{(2,0)} g = g$, which is usually written as $\psi_t^* g = g$. By (2.95), Killing fields always form a Lie algebra, whose associated Lie group (up to global analytic issues) is the subgroup of $\mathrm{Diff}(M)$ consisting of isometries.

In Minkowski space-time $(\mathbb{M}, \eta)$, the Christoffels symbols vanish (at least in the usual coordinates), so that $\nabla_\mu = \partial_\mu$ and $X_{\mu;\nu} = X_{\mu,\nu}$. Hence Killing fields satisfy $\partial_\mu X_\nu = -\partial_\nu X_\mu$, whose general solution is a 10-dimensional vector space (within $\mathfrak{X}(\mathbb{R}^4)$) with basis

$$\begin{aligned} X_{(\nu)} &= \partial_\nu; & X_{(\rho\sigma)} &= x_\rho\partial_\sigma - x_\sigma\partial_\rho; \\ X^\mu_{(\nu)}(x) &= \delta^\mu_\nu \ \ (\nu = 0,1,2,3); & X^\mu_{(\rho\sigma)}(x) &= x_\rho\delta^\mu_\sigma - x_\sigma\delta^\mu_\rho, \ \ (\rho,\sigma = 0,1,2,3),\end{aligned} \tag{3.75}$$

where $x_\rho = \eta_{\rho\sigma}x^\sigma$. This is the Lie algebra of the Poincaré-group (which is the subgroup of $GL_4(\mathbb{R})$ preserving the Minkowski metric η). See also Appendix A, §§A.1 - A.2.

[146] $\mathscr{L}_X$ is not a connection (as it fails to be $C^\infty(M)$-linear in X), but $\mathscr{L}_X$ and ∇_X both satisfy the Leibniz rule.

[147] Named after the German mathematician Wilhelm Killing (1847–1923), not the movie about Cambodja.

4 Curvature

The notion of curvature was originally introduces by Gauss in the context of lines in $\mathbb{R}^2$ and $\mathbb{R}^3$ and surfaces in $\mathbb{R}^3$. The modern approach via connections is highly abstract (and hence very powerful), but we shall recover at least some of the original ideas of Gauss c.s. later on.

4.1 Curvature tensor for general connections

For any connection ∇ on a vector bundle $E \to M$, the following map, indexed by $X, Y \in \mathfrak{X}(M)$,

$$\Omega(X,Y) : \Gamma(E) \to \Gamma(E); \tag{4.1}$$
$$\Omega(X,Y) := [\nabla_X, \nabla_Y] - \nabla_{[X,Y]} \tag{4.2}$$

is easily verified to be $C^\infty(M)$-linear in its argument $s \in \Gamma(E)$.[148] Furthermore, $\Omega(X,Y)$ is $C^\infty(M)$-linear in X and Y, so that we may equivalently write Ω as either of the following maps:

$$\tilde{\Omega} : \mathfrak{X}(M) \times \mathfrak{X}(M) \times \Gamma(E) \to \Gamma(E); \tag{4.3}$$
$$\hat{\Omega} : \mathfrak{X}(M) \times \mathfrak{X}(M) \times \Gamma(E^*) \times \Gamma(E) \to C^\infty(M), \tag{4.4}$$

where the first is three times $C^\infty(M)$-linear and the second four times so; the relationship between Ω as defined in (4.3) and $\hat{\Omega}$ is induced by a pointwise version of the (linear) isomorphism

$$\mathrm{Hom}(V^* \times V, \mathbb{R}) \cong \mathrm{Hom}(V,V); \tag{4.5}$$
$$\hat{\varphi}(\theta, v) = \theta(\varphi(v)). \tag{4.6}$$

In the usual basis (∂_μ) associated to a chart defining coordinates (x^μ) we may write (4.1) as

$$[\nabla_\mu, \nabla_\nu] s(x) = \Omega_{\mu\nu}(x) s(x), \tag{4.7}$$

where $\Omega_{\mu\nu} = \Omega(\partial_\mu, \partial_\nu)$ is a linear map $E_x \to E_x$. Relative to a local frame (u_a) for $\Gamma(E)$ in which $s(x) = s^\mathsf{a}(x) u_\mathsf{a}(x)$, with $s^\mathsf{a} \in C^\infty(U)$, see text after (3.56), we may therefore write

$$[\nabla_\mu, \nabla_\nu] s^\mathsf{a}(x) = \Omega^\mathsf{a}_{\mathsf{b}\mu\nu}(x) s^\mathsf{b}(x), \tag{4.8}$$

where, switching to the version (4.4), we have the coordinate- and basis-dependent expression

$$\Omega^\mathsf{a}_{\mathsf{b}\mu\nu} = \hat{\Omega}(\partial_\mu, \partial_\nu, e_\mathsf{b}, \omega^\mathsf{a}). \tag{4.9}$$

Thus the ***curvature tensor*** $\hat{\Omega}$ defined by a connection ∇ has four indices: the first two (i.e. a and b) refer to a basis of E_x, whereas the last two (viz. μ and ν) refer to a basis of $T_x M$.

In the case $E = TM$ the distinction between (μ, ν) and (a, b) is blurred. Our maps become

$$\Omega(X,Y) : \mathfrak{X}(M) \to \mathfrak{X}(M); \qquad Z \mapsto ([\nabla_X, \nabla_Y] - \nabla_{[X,Y]})Z; \tag{4.10}$$
$$\hat{\Omega} : \Omega(M) \times \mathfrak{X}(M) \times \mathfrak{X}(M) \times \mathfrak{X}(M) \to C^\infty(M); \qquad (\theta, Z, X, Y) \mapsto \theta(\Omega(X,Y)Z), \tag{4.11}$$

where in (4.11) we adopt the convention of moving $\Gamma(E^*) = \Omega(M)$ in (4.4) to the front.[149]

[148] It follows that $\Omega(X,Y)$ defines a cross-section of $\Gamma(\mathrm{Hom}(E,E))$.

[149] Regarding a connection as a map $\nabla : \Omega^p(E) \to \Omega^{p+1}(E)$, as in footnote 145, the corresponding curvature is simply defined as $\nabla^2 : \Omega^p(E) \to \Omega^{p+2}(E)$, so that $\nabla^2 u = R \wedge u$ for some $R \in \Omega^2(E)$.

4.2 Riemann tensor

We now fix a metric on M and take ∇ to be the Levi-Civita connection on TM defined by g. Denoting $\hat{\Omega}$ by Riem (or R), we obtain the ***Riemann tensor*** $\mathrm{Riem} \in \mathfrak{X}^{(3,1)}(M)$, defined by

$$\mathrm{Riem}(\theta,Z,X,Y) := \theta(\Omega(X,Y)Z) = \theta([\nabla_X,\nabla_Y]-\nabla_{[X,Y]})Z). \tag{4.12}$$

In coordinates, where $R^\rho_{\sigma\mu\nu} = \mathrm{Riem}(dx^\rho,\partial_\sigma,\partial_\mu,\partial_\nu)$ and $[\partial_\mu,\partial_\nu]=0$, we therefore have

$$[\nabla_\mu,\nabla_\nu]Z^\rho = R^\rho_{\sigma\mu\nu}Z^\sigma; \tag{4.13}$$

$$R^\rho_{\sigma\mu\nu} = \Gamma^\rho_{\sigma\nu,\mu} - \Gamma^\rho_{\sigma\mu,\nu} + \Gamma^\rho_{\mu\tau}\Gamma^\tau_{\nu\sigma} - \Gamma^\rho_{\nu\tau}\Gamma^\tau_{\mu\sigma}, \tag{4.14}$$

where the Christoffel symbols are defined by (3.25), i.e., this time in Greek indices,

$$\Gamma^\rho_{\mu\nu} = \tfrac{1}{2}g^{\rho\sigma}(g_{\sigma\mu,\nu}+g_{\sigma\nu,\mu}-g_{\mu\nu,\sigma}). \tag{4.15}$$

A (semi) Riemannian manifold (M,g) is ***locally flat*** (or locally isometric to a flat space) if each point $x \in M$ has a coordinate nbhd U with a chart $\varphi : U \to \mathbb{R}^n$ and associated coordinates $x^\mu = \varphi^\mu(x)$, see Definition 2.1.2, in which the metric is flat, i.e. $g_{\mu\nu}(x) = \delta_{\mu\nu}$ for each $x \in U$ in the Riemannian case, $g_{\mu\nu}(x) = \eta_{\mu\nu}$ in the Lorentzian case, etc. The first nontrivial result about the Riemann tensor (which was known to Riemann himself) is that it detects local flatness:

Theorem 4.1 *A (semi) Riemannian manifold (M,g) is locally flat iff* Riem $= 0$*, that is,*

$$R^\rho_{\sigma\mu\nu} = 0. \tag{4.16}$$

One direction is trivial: if $g_{\mu\nu}(x) = \delta_{\mu\nu}$ (etc.), then the Christoffel symbols (4.15) vanish, so that (4.14) vanishes. Proving local flatness from $R^\rho_{\sigma\mu\nu} = 0$ relies on the ***Frobenius theorem***:[150]

Lemma 4.2 *If* Riem $= 0$*, i.e.* (4.16)*, then each $x \in M$ has an open nbhd U such that for any $v \in T_xM$ there is a unique vector field $Z \in \mathfrak{X}(U)$ with $Z(x) = v$ and $\nabla_X Z = 0$ for all $X \in \mathfrak{X}(U)$.*

Proof. We just sketch the proof and explain the role of (4.16). In local coordinates (x^μ) the condition $\nabla_X Z = 0$ for all X is equivalent to $\nabla_\mu Z^\rho = 0$ for all μ. One can solve

$$\nabla_\mu Z^\rho(x^1,\ldots,x^n) = 0; \qquad Z^\rho(x_0) = v^\rho, \tag{4.17}$$

first for $\mu = 1$ at fixed $(x^2,\ldots,x^n)$, then for $\mu = 2$ at fixed $(x^1,x^3,\ldots,x^n)$, etc. The integrability condition $[\nabla_\mu,\nabla_\nu]Z^\rho = 0$ for this procedure is satisfied, since by (4.13), this is the same as $R^\rho_{\sigma\mu\nu}Z^\sigma = 0$, which holds by assumption, as in (4.16). □

The thrust of the Frobenius theorem, then, is that the *necessary* condition (4.16) for the solution of all equations $\nabla_X Z = 0$ is also *sufficient*. To prove the nontrivial direction of Theorem 4.1, take an orthonormal basis (e_a) of T_xM (which exists because g_x can be diagonalized at any point $x \in M$) and extend this to a frame $(e_a(y))$ defined for each $y \in U$ (as in Lemma 4.2), so that

$$e_a(x) = e_a \qquad \nabla_X e_a = 0, \tag{4.18}$$

[150] This holds for any vector bundle $E \to M$ with connection ∇: if $\Omega = 0$, then each $x \in M$ has an open nbhd U such that for any $v \in E_x$ there is a unique local section $s \in \Gamma(U,E)$ with $s(x) = v$ and $\nabla_X s = 0$ for all $X \in \mathfrak{X}(U)$. In this generalised version the lemma is proved in e.g. Heckman (2017), Theorem 2.34

for all X. Property (3.50) of the Levi-Civita connection then gives $X(g(e_a,e_b)) = 0$. Hence

$$g_y(e_a,e_b) = g_x(e_a,e_b) = \delta_{ab}, \tag{4.19}$$

and similarly for other signatures, for all $y \in U$. In particular, the vectors (e_a) remain orthonormal throughout U and hence remain a basis of each T_yM, $y \in U$. Moreover, since ∇ is torsion-free,

$$[e_a,e_b] = \nabla_{e_a}e_b - \nabla_{e_b}e_a = 0 - 0 = 0. \tag{4.20}$$

Using (2.34) and (2.35), this can be used to show that the flows of all vector fields e_a commute, which in turn implies that there is an open subset $V_x \subset T_xM$ such that the map

$$t^a e_a(x) \mapsto \varphi^{(1)}_{t^1} \circ \cdots \circ \varphi^{(n)}_{t^n}(x), \tag{4.21}$$

where $\varphi^{(a)}_t$ is the flow of the vector field e_a emanating from x (i.e., with initial value $\varphi^{(a)}_0 = x$), is a diffeomorphism from V_x onto its image $U' \subset U$ in M. If the image point of $t^a e_a$ under this map is y, we then define its coordinates to be $(y^a = t^a)$.[151] By construction, $e_a = \partial/\partial y^a$, so that

$$g_y(\partial_a,\partial_b) = g_y(e_a,e_b) = \delta_{ab}. \qquad \square$$

Proposition 4.3 *Any torsion-free connection satisfies the* **Bianchi identities**:[152]

$$\Omega(X,Y)Z + \Omega(Y,Z)X + \Omega(Z,X)Y = 0; \tag{4.22}$$
$$(\nabla_X\Omega)(Y,Z) + (\nabla_Y\Omega)(Z,X) + (\nabla_Z\Omega)(X,Y) = 0. \tag{4.23}$$

For the Levi-Civita connection, these identities read

$$R^\rho_{\sigma\mu\nu} + R^\rho_{\mu\nu\sigma} + R^\rho_{\nu\sigma\mu} = 0; \tag{4.24}$$
$$R^\rho_{\sigma\mu\nu;\tau} + R^\rho_{\sigma\tau\mu;\nu} + R^\rho_{\sigma\nu\tau;\mu} = 0. \tag{4.25}$$

Proof. The first one, in the form (4.22) using the definition (4.2), is most simply proved by taking commuting vector-fields X, Y, and Z, such as, in coordinates, $X = \partial_\mu$, $Y = \partial_\nu$, $Z = \partial_\sigma$, which indeed leads to (4.24). One then finds that $\Omega(X,Y)Z + \Omega(Y,Z)X + \Omega(Z,X)Y$ is equal to

$$\nabla_X\nabla_Y Z - \nabla_Y\nabla_X Z + \nabla_Y\nabla_Z X - \nabla_Z\nabla_Y Z + \nabla_Z\nabla_X Y - \nabla_X\nabla_Z Y,$$

which vanishes if torsion-freeness (3.47) is taken into account, which means $\nabla_X Y = \nabla_Y X$.

The second one is usually proved by using geodesic normal coordinates, cf. §5.1. Assuming the reader is familiar with these, at the origin of these coordinates the Riemann tensor equals

$$R^\rho_{\sigma\mu\nu} = \tfrac{1}{2}g^{\rho\tau}(\partial_\sigma\partial_\mu g_{\nu\tau} - \partial_\nu\partial_\sigma g_{\mu\tau} + \partial_\nu\partial_\tau g_{\mu\sigma} - \partial_\mu\partial_\tau g_{\nu\sigma}). \tag{4.26}$$

Since at the origin $\nabla_\tau R^\rho_{\sigma\mu\nu} = \partial_\tau R^\rho_{\sigma\mu\nu}$, where ∇_τ can even be taken inside the brackets in (4.26), the identity (4.25) easily follows.[153] $\square$

The real nature of both Bianchi identities is that they are a consequence of the covariance property

[151] This construction defines geodesic normal coordinates in the special case at hand, as will be seen in §5.2.
[152] Continuing footnote 149, The differential Bianchi identity (4.23) simply reads $\nabla R = 0$.
[153] Another, more abstract proof of (4.23) follows from Cartan's exterior calculus and the previous footnote.

$$\psi^*_{(3,1)}\mathrm{Riem}_g = \mathrm{Riem}_{\psi^*_{(2,0)}g}, \tag{4.27}$$

or $\psi^*\mathrm{Riem}_g = \mathrm{Riem}_{\psi^* g}$, for any diffeomorphism ψ. Here $\psi^*_{(k,l)}$ is defined as in (2.86), and we have indicated the dependence of the Riemann tensor on the metric. First, eq. (4.27) reads

$$\mathrm{Riem}_g((\psi^{-1})^*\theta, \psi_* Z, \psi_* X, \psi_* Y) = \mathrm{Riem}_{\psi^* g}(\theta, Z, X, Y). \tag{4.28}$$

Using the definition (4.12) of the Riemann tensor, this follows from the underlying property

$$\nabla^{\psi^* g}_X Y = \psi_*^{-1}(\nabla^g_{\psi_* X}(\psi_* Z)), \tag{4.29}$$

where ∇^g is the Levi-Civita connection for the metric g. Eq. (4.29), in turn, follows from Theorem 3.9, notably from the uniqueness of connections satisfying

$$X(\psi^* g(Y,Z)) = \psi^* g(\nabla^{\psi^* g}_X Y, Z) + \psi^* g(Y, \nabla^{\psi^* g}_X Z). \tag{4.30}$$

To see this, one defines a connection ∇' by $\nabla'_X Y = \psi_*^{-1}(\nabla^g_{\psi_* X}(\psi_* Z))$ and shows that

$$X(\psi^* g(Y,Z)) = \psi^* g(\nabla'_X Y, Z) + \psi^* g(Y, \nabla'_X Z). \tag{4.31}$$

Eq. (4.27) is also true for a one-parameter family ψ_s, i.e. $\psi_s^*\mathrm{Riem}_g = \mathrm{Riem}_{\psi_s^* g}$; taking $d\psi_s/ds$ at $s=0$ yields both Bianchi identities.[154] The left-hand side equals the Lie derivative $\mathscr{L}_X \mathrm{Riem}_g$, where X is the vector field whose flow is ψ_s, and this may in term be expressed in terms of the covariant derivatives of Riem_g using (3.72). The right-hand side may be computed uses the techniques explained in §7.2, notably (7.31) and its consequences for Riem and (7.47) - (7.49). After lengthy calculations and multiple cancellations, one finds that both derivatives are equal iff

$$X^\tau(R^\rho_{\sigma\mu\nu;\tau} + R^\rho_{\sigma\tau\mu;\nu} + R^\rho_{\sigma\nu\tau;\mu}) = \tfrac{1}{2}(\nabla_\mu(X^\tau B^\rho_{\nu\tau\sigma}) - \nabla_\nu(X^\tau B^\rho_{\mu\tau\sigma})), \tag{4.32}$$

where $B^\rho_{\sigma\mu\nu} = R^\rho_{\sigma\mu\nu} + R^\rho_{\mu\nu\sigma} + R^\rho_{\nu\sigma\mu}$, cf. (4.24). Choosing $X=0$ at some given point gives

$$(\nabla_\mu X^\tau) B^\rho_{\nu\tau\sigma} = (\nabla_\nu X^\tau) B^\rho_{\mu\tau\sigma}. \tag{4.33}$$

Taking $\nabla_\mu X^\tau = \delta^\tau_\mu$ and using $B^\rho_{\nu\tau\sigma} = -B^\rho_{\tau\nu\sigma}$ inherited from $R^\rho_{\sigma\nu\mu} = -R^\rho_{\sigma\mu\nu}$, cf. (4.13), then forces $B^\rho_{\nu\tau\sigma} = 0$, i.e. (4.24). Putting this in (4.32) and choosing $X^\tau \neq 0$ then gives (4.25).

We can lower the first index of the Riemann tensor to obtain $\mathrm{Riem}^\flat \in \mathfrak{X}^{(4,0)}(M)$, that is,

$$\mathrm{Riem}^\flat(W,Z,X,Y) = g(W, (\Omega(X,Y)Z)) = g(W, ([\nabla_X, \nabla_Y] - \nabla_{[X,Y]})Z); \tag{4.34}$$

$$R^\flat_{\rho\sigma\mu\nu} = g_{\rho\tau} R^\tau_{\sigma\mu\nu} \equiv R_{\rho\sigma\mu\nu}. \tag{4.35}$$

We omit the "flat" suffix. This leads to some more identities satisfied by the Riemann tensor:

$$R_{\rho\sigma\nu\mu} = -R_{\rho\sigma\mu\nu}; \tag{4.36}$$

$$R_{\sigma\rho\mu\nu} = -R_{\rho\sigma\mu\nu}; \tag{4.37}$$

$$R_{\mu\nu\rho\sigma} = R_{\rho\sigma\mu\nu}, \tag{4.38}$$

of which the first is trivial from (4.13) and hence did not require lowering indices, the second states that each map $\Omega(X,Y)$ is an isometry of $T_x M$, and the third is conceptually bizarre, since, as we explained, the first pair of indices plays a completely different role from the second (and yet one can apparently interchange them). Its proof is straightforward from (4.14) - (4.15), but to avoid a long calculation one may again want to use using geodesic normal coordinates, in which, from (4.26), at the origin one has an expression rapidly yielding (4.38), namely

$$R_{\rho\sigma\mu\nu} = \tfrac{1}{2}(\partial_\sigma\partial_\mu g_{\nu\rho} - \partial_\nu\partial_\sigma g_{\mu\rho} + \partial_\nu\partial_\rho g_{\mu\sigma} - \partial_\mu\partial_\rho g_{\nu\sigma}). \tag{4.39}$$

[154] See Kazdan (1981). Einstein's contracted Bianchi identity (7.56) will be proved separately in §7.2.

4.3 Sectional curvature and *Theorema Egregium*

All information in the Riemann tensor is in the so-called *sectional curvature*. Here is the key:[155]

Proposition 4.4 *The (pointwise) Riemann tensor* $\mathrm{Riem}_x \in (T_x^*M)^{\otimes 4}$ *is equivalent to a map*

$$\widehat{\mathrm{Riem}}_x : \Lambda^2 T_xM \to \Lambda^2 T_xM, \tag{4.40}$$

which is linear and self-adjoint (i.e. symmetric) with respect to the inner product

$$\langle X_1 \wedge X_2, Y_1 \wedge Y_2 \rangle_x := g_x(X_1,Y_1)g_x(X_2,Y_2), \tag{4.41}$$

where $X \wedge Y := \frac{1}{2}(X \otimes Y - Y \otimes X)$. *Thus* Riem_x *is specified by the associated quadratic form*

$$Q_x : \Lambda^2 T_xM \to \mathbb{R}; \qquad X \wedge Y \mapsto \langle X \wedge Y, \widehat{\mathrm{Riem}}_x(X \wedge Y)\rangle_x = \mathrm{Riem}_x(X,Y,X,Y). \tag{4.42}$$

Proof. We first show that Riem_x is equivalent to a linear map

$$\widetilde{\mathrm{Riem}}_x : T_xM \otimes T_xM \to T_xM \otimes T_xM. \tag{4.43}$$

1. Recalling (4.1) - (4.2) and (4.10), we have $\Omega_x(X,Y) \in \mathrm{Hom}(T_xM, T_xM)$ by definition.
2. Linear extension of $\theta \otimes v \mapsto (w \mapsto \theta(w)v$ gives an isomorphism $V^* \otimes V \xrightarrow{\cong} \mathrm{Hom}(V,V)$.
3. A metric on V gives $V^* \cong V$ canonically (cf. §2.3), so that $\mathrm{Hom}(V,V) \cong V \otimes V$.

By the symmetry (4.38), the map (4.43) is self-adjoint with respect to the bilinear form

$$\langle X_1 \otimes X_2, Y_1 \otimes Y_2 \rangle_x = g_x(X_1,Y_1)g_x(X_2,Y_2). \tag{4.44}$$

Because of the symmetries (4.36) - (4.38), both the map $\widetilde{\mathrm{Riem}}_x$ and the bilinear form (4.44) restrict to the linear subspace $\Lambda^2 T_xM \subset T_xM \otimes T_xM$, without any loss of information. □

Explicitly, the map (4.40) is given by linear extension of

$$\widehat{\mathrm{Riem}}_x : \partial_\mu \wedge \partial_\nu \mapsto g^{\alpha\sigma} R^\rho_{\sigma\mu\nu} \partial_\rho \wedge \partial_\alpha. \tag{4.45}$$

It is easy to show that $X, Y \in T_xM$ are linearly independent iff $P_x(X \wedge Y) \neq 0$, where

$$P_x(X \wedge Y) := g_x(X \wedge Y, X \wedge Y) = g_x(X,X)g_x(Y,Y) - g_x(X,Y)^2 \tag{4.46}$$

is the square of the (metric) area of the parallelogram in T_xM with sides X and Y, up to a sign.

Definition 4.5 *If* $P_x(X,Y) \neq 0$, *the* **sectional curvature** $C_x(X \wedge Y)$ *of the X-Y plane is given by*

$$C_x(X \wedge Y) := \frac{Q_x(X \wedge Y)}{P_x(X \wedge Y)} = \frac{\mathrm{Riem}_x(X,Y,X,Y)}{g_x(X,X)g_x(Y,Y) - g_x(X,Y)^2}. \tag{4.47}$$

[155] Let V be a (real) vector space. Defining $\tau : V \otimes V \to V \otimes V$ by linear extension of $v \otimes w \mapsto w \otimes v$, the space $\Lambda^2 V \equiv V \otimes_A V \subset V \otimes V$ is the antisymmetric part of $V \otimes V$, defined as the eigenspace of τ with eigenvalue -1. Furthermore, if $T : W \to W$ is linear and symmetric with respect to some inner product $\langle \cdot, \cdot \rangle$ on W, i.e., $\langle X, TY \rangle = \langle TX, Y \rangle$ for all $X, Y \in W$, then the associated quadratic form $Q : W \to \mathbb{R}$ is defined by $Q(X) = \langle X, TX \rangle$. The map T may be then be recovered from Q (and the inner product) via the formula $\langle X, TY \rangle = \frac{1}{4}(Q(X+Y) - Q(X-Y))$.

The specific combination in (4.47) makes $C_x(X \wedge Y)$ independent of the choice of X and Y within the plane (in T_xM) they span, *and hence makes C_x a function of that plane only*. Moreover, Proposition 4.4 shows that we may interchangeably use either the Riemann tensor itself or its associated sectional curvatures. For an orthonormal pair $X = e_a$, $Y = e_b$ we simply have

$$C_x(e_a, e_b) = \mathrm{Riem}_x(e_a, e_b, e_a, e_b). \tag{4.48}$$

We now explain how the notion of sectional curvature is related to the classical differential geometry of surfaces, especially through the famous *Theorema Egregium* of Gauss from 1828.

The classical theory of surfaces $\Sigma \subset \mathbb{R}^3$ was largely based on local constructions. Let $U \subset \mathbb{R}^2$ be open and let $F : U \to \mathbb{R}^3$ be a smooth map that is a homeomorphism onto its image $F(S) = \Sigma$ and also has injective derivatives $F'_u : T_uS \to T_{F(u)}M$ for all $u \in S$ (equivalently, F'_u has rank 2).

If $u = (u^1, u^2)$ are the standard coordinates on U, we simply say $F(u^1, u^2) \in \Sigma \subset \mathbb{R}^3$ has coordinates (u^2, u^2), too. This gives three canonical vector fields in $\mathbb{R}^3$ defined on Σ, viz.[156]

$$\vec{x}_1 := F'(\partial/\partial u^1); \qquad \vec{x}_2 := F'(\partial/\partial u^2); \qquad \vec{N} := \vec{x}_1 \times \vec{x}_2 / \|\vec{x}_1 \times \vec{x}_2\|. \tag{4.49}$$

The vectors $\vec{x}_1$ and $\vec{x}_2$ are tangent to Σ, whereas $\vec{N}$ is orthogonal to Σ. Since the pair $(\vec{x}_1, \vec{x}_2)$ is a basis of $T_{F(u)}\Sigma$, $u \in U$, the triple $(\vec{x}_1, \vec{x}_2, \vec{N})$ is a basis of $T_u\mathbb{R}^3 \cong \mathbb{R}^3$. Early Greek alphabet indices α, β etc. run through 1, 2, whereas $i, j, k = 1, 2, 3$. The following two tensors on Σ go back to Gauss (and will be used in a similar way in the PDE approach to GR, see chapter 8):

1. The ***first fundamental form*** $\tilde{g}$ is the metric induced by the Euclidean metric δ on $\mathbb{R}^3$, i.e.

$$\tilde{g}_{\alpha\beta} = \tilde{g}(\partial_\alpha, \partial_\beta) = \langle \vec{x}_\alpha, \vec{x}_\beta \rangle = \sum_{i=1}^{3} \frac{\partial F^i}{\partial u^\alpha} \cdot \frac{\partial F^i}{\partial u^\beta}. \tag{4.50}$$

   Note that although the (∂_1, ∂_2) basis is orthonormal in $U \subset \mathbb{R}^2$, its pushforward $(\vec{x}_1, \vec{x}_2)$ to Σ may no longer be orthonormal in $\mathbb{R}^3$: this depends on the embedding map F.

2. The ***second fundamental form*** or ***extrinsic curvature*** (a more telling name!) $\tilde{k}$ of the embedding, is constructed as follows. First, for $X = X^\alpha \vec{x}_\alpha \in \mathfrak{X}(\Sigma)$ we define the 3-vector

$$\nabla_X \vec{N} = X^\alpha \frac{\partial \vec{N}}{\partial u^\alpha}. \tag{4.51}$$

   If $X_u \equiv X_{F(u)}$ is tangent to a curve $F(\gamma^1(t), \gamma^2(t))$, then $X^\alpha = d\gamma^\alpha/dt_{|t=0}$. We may then also write $\nabla_X \vec{N}(u, v) = d\vec{N}(\gamma^1(t), \gamma^2(t))/dt_{|t=0}$ (the notation ∇_X is used because from a "higher perspective" one uses covariant differentiation with respect to the Levi-Civita connection defined by the flat metric δ on $\mathbb{R}^3$). One could also simply say that

$$\nabla_X N^i = X(N^i) = X^\alpha \partial_\alpha N^i \qquad (i = 1, 2, 3), \tag{4.52}$$

   which is (3.36) with vanishing Christoffel symbols (in $\mathbb{R}^3$). Since $\langle \vec{N}, \vec{N} \rangle = 1$, we have

$$0 = X(1_\Sigma) = X(\langle \vec{N}, \vec{N} \rangle) = \langle \nabla_X \vec{N}, \vec{N} \rangle + \langle \vec{N}, \nabla_X \vec{N} \rangle = 2\langle \nabla_X \vec{N}, \vec{N} \rangle, \tag{4.53}$$

   so that $\nabla_X \vec{N}$ is orthogonal to $\vec{N}$ (in $\mathbb{R}^3$), and hence it must be tangent to Σ. This gives rise to the ***Weingarten map*** (with a conventional minus sign for historical reasons)

$$W : T\Sigma \to T\Sigma; \qquad X \mapsto -\nabla_X \vec{N}. \tag{4.54}$$

[156] Injectivity of F' implies that the denominator in (4.49) is nonzero.

In terms of the Weingarten map, Gauss and Monge defined two curvature *scalars*, namely

$$K = \det(W) = \kappa_1 \kappa_2 \qquad \textbf{\textit{(Gauss curvature)}}; \tag{4.55}$$

$$H = \mathrm{tr}(W) = \kappa_1 + \kappa_2 \qquad \textbf{\textit{(mean curvature)}}, \tag{4.56}$$

where κ_1 and κ_2 are the eigenvalues of W, as well as the extrinsic curvature *tensor*, i.e. $\tilde{k}$,

$$\tilde{k}(X,Y) := \tilde{g}(W(X),Y) = -\tilde{g}(\nabla_X \vec{N}, Y) = -\langle \nabla_X \vec{N}, Y \rangle. \tag{4.57}$$

It is easy to show that the extrinsic curvature tensor thus defined is *symmetric*, i.e.,

$$\tilde{k}(X,Y) = \tilde{k}(Y,X), \tag{4.58}$$

which is the same as $\langle \nabla_Y \vec{N}, X \rangle = \langle \nabla_X \vec{N}, Y \rangle$. To see this, note that $\langle \vec{N}, X \rangle = 0$ (since X and Y are tangent to Σ and hence orthogonal to $\vec{N}$), hence $0 = Y(\langle \vec{N}, X \rangle) = \langle \nabla_Y \vec{N}, X \rangle + \langle \vec{N}, \nabla_Y X \rangle$. Since ∇ (as the flat Levi-Civita connection on $\mathbb{R}^3$) is torsion-free, we have $\nabla_Y X = \nabla_X Y - [X,Y]$, so

$$\langle \nabla_Y \vec{N}, X \rangle = -\langle \vec{N}, \nabla_Y X \rangle = -\langle \vec{N}, \nabla_X Y \rangle + \langle \vec{N}, [X,Y] \rangle = -\langle \vec{N}, \nabla_X Y \rangle = \langle \nabla_X \vec{N}, Y \rangle. \tag{4.59}$$

Here we also used $\langle \vec{N}, [X,Y] \rangle = 0$, because $[X,Y]$ is tangent to Σ whenever X and Y are. This computation also yields an alterative expression for $\tilde{k}$, which is manifestly symmetric:

$$\tilde{k}_{\alpha\beta} = \langle \vec{x}_{\alpha\beta}, \vec{N} \rangle; \qquad \vec{x}_{\alpha\beta} \equiv \partial_\beta \vec{x}_\alpha, \tag{4.60}$$

where, in terms of $F : U \to \mathbb{R}^3$, the components $x^i_{\alpha\beta}$ of the vector $\vec{x}_{\alpha\beta}$ are simply given by

$$x^i_{\alpha\beta} = \frac{\partial^2 F^i}{\partial u^\alpha \partial u^\beta}. \tag{4.61}$$

The relationship between the two curvature scalars and the two fundamental forms is

$$K = \det(\tilde{k}) / \det(\tilde{g}); \tag{4.62}$$

$$H = \mathrm{tr}(\tilde{g}^{-1}\tilde{k}) = \sum_{i,j=1,2} \tilde{g}^{ij} \tilde{k}_{ij}. \tag{4.63}$$

These objects are very useful, if only because they are quite easy to compute in practice:

- In the simplest case (from which all others follow by translation and rotation), a ***plane*** in $\mathbb{R}^3$ is parametrized by $(x = u^1, y = u^2, z = 0)$, i.e., officially,

  $$F^1(u^1,u^2) = u^1; \qquad F^2(u^1,u^2) = u^2; \qquad F^3(u^1,u^2) = 0. \tag{4.64}$$

  The induced metric follows from (4.50) as $\tilde{g}_{11} = \tilde{g}_{22} = 1$, i.e.

  $$\tilde{g} = (du^1)^2 + (du^2)^2. \tag{4.65}$$

  Eq. (4.49) gives the normal as $\vec{N} = (0,0,1)$, which is independent of (u^1,u^2), so

  $$\tilde{k} = 0. \tag{4.66}$$

  Consequently, $H = K = 0$, as expected.

- The ***cylinder*** C_a with radius a is defined by $x^2+y^2=a^2$ and z arbitrary, and hence may be parametrized by $(u^1,u^2)=(\varphi,z)$, where $\varphi\in[0,2\pi]$ and $z\in\mathbb{R}$, so that

$$(x=a\cos u^1, y=a\sin u^1, z=u^2). \tag{4.67}$$

This time the induced metric is

$$\tilde{g}=a^2d\varphi^2+dz^2, \tag{4.68}$$

whereas the normal $\vec{N}(\varphi,z)=(\cos\varphi,\sin\varphi,0)$ leads to

$$\tilde{k}=-ad\varphi^2. \tag{4.69}$$

Since $\tilde{g}^{11}\equiv\tilde{g}^{\varphi\varphi}=1/a^2$, this gives

$$K=0; \qquad\qquad H=-\frac{1}{a}. \tag{4.70}$$

This is a very natural result: the larger a is, the more the cylinder *locally* approximates a plane, whose extrinsic curvature vanishes.

- Finally, the ***sphere*** S^2_a is defined by $x^2+y^2+z^2=a^2$ and hence we may define

$$x=a\sin\theta\cos\varphi; \qquad y=a\sin\theta\sin\varphi; \qquad z=a\cos\theta, \tag{4.71}$$

which of course gives the well-known "round" metric

$$\tilde{g}=a^2d\Omega; \qquad\qquad d\Omega:=d\theta^2+\sin^2\theta d\varphi^2. \tag{4.72}$$

The normal vector is $\vec{N}(\theta,\varphi)=(\sin\theta\cos\varphi,\sin\theta\sin\varphi,\cos\theta)$, which gives

$$\tilde{k}=-a^{-1}\tilde{g}=-a(d\theta^2+\sin^2\theta d\varphi^2), \tag{4.73}$$

so that

$$K=\frac{1}{a^2}; \qquad\qquad H=-\frac{2}{a}. \tag{4.74}$$

Somewhat anachronistically compared to Gauss, we now define the Riemann tensor $\tilde{R}^{\delta}_{\alpha\gamma\beta}$ in terms of the metric $\tilde{g}$ on Σ as in (4.14), and lower the first index with $\tilde{g}$ as usual.[157] Since Σ is two-dimensional, there is just one sectional curvature, given, from (4.47), by

$$C=C(\vec{x}_1,\vec{x}_2)=R_{1212}/\det(\tilde{g}). \tag{4.75}$$

In slightly modernized form, then, the ***Theorema Egregium*** of Gauss states that

$$K=C. \tag{4.76}$$

Gauss found this theorem remarkable because it equates K, which is *a priori* defined *extrinisically* through the Weingarten map W and hence through the embedding of Σ in $\mathbb{R}^3$, with C, which is defined via the *intrinsic* geometry of Σ as encoded by its internal metric $\tilde{g}_{\alpha\beta}$.

[157] This, as well as index raising, applies generally to all tensors on Σ, e.g. $\tilde{k}^{\delta}_{\gamma}=\tilde{g}^{\delta\beta}\tilde{k}_{\beta\gamma}$. Also, $\tilde{k}_{\alpha\gamma,\beta}=\partial_\beta\tilde{k}_{\alpha\gamma}$.

The proof relies on equations that are of independent interest (and will resurface in GR):

$$\vec{x}_{\alpha\beta} = \tilde{\Gamma}^{\gamma}_{\alpha\beta}\vec{x}_{\gamma} + \tilde{k}_{\alpha\beta}\vec{N}; \qquad \textbf{\textit{(Gauss)}} \tag{4.77}$$

$$\partial_{\alpha}\vec{N} = -\tilde{k}^{\beta}_{\alpha}\vec{x}_{\beta}; \qquad \textbf{\textit{(Weingarten)}} \tag{4.78}$$

$$\tilde{R}^{\delta}_{\alpha\gamma\beta} = \tilde{k}^{\delta}_{\gamma}\tilde{k}_{\alpha\beta} - \tilde{k}^{\delta}_{\beta}\tilde{k}_{\alpha\gamma}; \qquad \textbf{\textit{(Gauss)}} \tag{4.79}$$

$$\tilde{k}_{\alpha\beta,\gamma} + \tilde{\Gamma}^{\delta}_{\alpha\beta}\tilde{k}_{\gamma\delta} = \tilde{k}_{\alpha\gamma,\beta} + \tilde{\Gamma}^{\delta}_{\alpha\gamma}\tilde{k}_{\beta\delta}, \qquad \textbf{\textit{(Codazzi)}} \tag{4.80}$$

where the $\tilde{\Gamma}^{\gamma}_{\alpha\beta}$ are the Christoffel symbols (as originally introduced!) associated to the metric $\tilde{g}$ on Σ, and $\tilde{k}^{\beta}_{\alpha} = \tilde{g}^{\beta\gamma}\tilde{k}_{\alpha\gamma}$, where $(\tilde{g}^{\beta\gamma})$ is the inverse matrix to $(\tilde{g}_{\beta\gamma})$, as usual. Weingarten's eq. (4.78) is just a restatement of (4.57), and hence is the definition of $\tilde{k}_{\alpha\beta}$. Gauss's eq. (4.77) is simply the expansion of the 3-vectors $\vec{x}_{\alpha\beta}$ in terms of the basis $(\vec{x}_u, \vec{x}_v, \vec{N})$. The specific form $\tilde{k}_{\alpha\beta}$ of the coefficient of $\vec{N}$ immediately follows from (4.60). To derive the coefficient of $\vec{x}_{\gamma}$, let us assume (4.77) for initially unknown coefficients $\tilde{\Gamma}^{\gamma}_{\alpha\beta}$. We then obtain

$$\langle \vec{x}_{\gamma}, \vec{x}_{\alpha\beta} \rangle = \tilde{\Gamma}^{\delta}_{\alpha\beta}\langle \vec{x}_{\gamma}, \vec{x}_{\delta} \rangle = \tilde{g}_{\gamma\delta}\tilde{\Gamma}^{\delta}_{\alpha\beta}, \tag{4.81}$$

so that $\tilde{\Gamma}^{\gamma}_{\alpha\beta} = g^{\gamma\delta}\langle \vec{x}_{\delta}, \vec{x}_{\alpha\beta} \rangle$. The relation (3.25) then follows from (4.60), which yields

$$2\langle \vec{x}_{\delta}, \vec{x}_{\alpha\beta} \rangle = \partial_{\beta}\langle \vec{x}_{\delta}, \vec{x}_{\alpha} \rangle + \partial_{\alpha}\langle \vec{x}_{\delta}, \vec{x}_{\beta} \rangle - \partial_{\delta}\langle \vec{x}_{\alpha}, \vec{x}_{\beta} \rangle. \tag{4.82}$$

The Gauss–Codazzi equations (4.79) - (4.80) then follow from the integrability condition

$$\partial_{\gamma}\partial_{\beta}\vec{x}_{\alpha} = \partial_{\beta}\partial_{\gamma}\vec{x}_{\alpha}, \tag{4.83}$$

i.e., $\partial_{\gamma}\vec{x}_{\alpha\beta} = \partial_{\beta}\vec{x}_{\alpha\gamma}$. Indeed, the Gauss–Weingarten equations (4.77) - (4.78) give

$$\vec{x}_{\alpha\beta\gamma} - \vec{x}_{\alpha\gamma\beta} = (\tilde{R}^{\delta}_{\alpha\gamma\beta} - \tilde{k}^{\delta}_{\gamma}\tilde{k}_{\alpha\beta} + \tilde{k}^{\delta}_{\beta}\tilde{k}_{\alpha\gamma})\vec{x}_{\delta} + (\tilde{k}_{\alpha\beta,\gamma} + \tilde{\Gamma}^{\delta}_{\alpha\beta}\tilde{k}_{\gamma\delta} - \tilde{k}_{\alpha\gamma,\beta} + \tilde{\Gamma}^{\delta}_{\alpha\gamma}\tilde{k}_{\beta\delta})\vec{N}, \tag{4.84}$$

so that Gauss's equation (4.79) is the component of (4.83) tangential (to Σ), whilst Codazzi's equation (4.80) is its normal component. The *Theorema Egregium* now follows from (4.79), since (4.76) is the same as $\det(\tilde{k}) = R_{1212}$. □

Take the cylinder, whose metric (4.68) is flat. Hence (4.76) is just $0 = 0$ (and this is of course also true for the plane). The sphere is less trivial; either direct computation or eq. (4.85) and Theorem 4.8 in the next section show that $R_{1212} = g_{11}g_{22}/a^2 = a^2\sin^2\theta$, so that, with $\det(\tilde{g}) = a^4\sin^2\theta$, we find $R_{1212}/\det(\tilde{g}) = 1/a^2$, which, given (4.74), confirms (4.75) - (4.76).

Finally, we return to the interpretation of sectional curvature in general (semi) Riemannian geometry. In §5.2 we will see that each $x \in M$ has a so-called ***normal neighbourhood*** U_x that is diffeomorphic to some subspace $\mathscr{V}_x \subset T_xM$ through the ***exponential map*** $\exp_x : \mathscr{V}_x \to M$. Take linearly independent vectors $X, Y \in T_xM$ with associated plane $\mathrm{span}(X,Y) \subset T_xM$, and consider the two-dimensional submanifold $\Sigma_{X,Y} = \exp_x(\mathrm{span}(X,Y) \cap \mathscr{V}_x) \subset U_x$ of M; note that $\Sigma_{X,Y}$ is spanned by geodesics emanating from x that have tangent vectors in $\mathrm{span}(X,Y)$. This surface has an intrinsically defined Gaussian curvature K, which, at x, by the *Theorema Egregium* is just its sectional curvature $C_x(X,Y)$. It follows that, through its associated sectional curvatures (which in turn define it), the Riemann tensor gives the Gaussian curvatures K of all possible two-dimensional subspaces of M. Conversely, these quantities give a complete description of the Riemann tensor. Its original definition (4.12) through the covariant derivative, which is very abstract, therefore has an interpretation in classical two-dimensional differential geometry.

4.4 Spaces of constant curvature

As another take on sectional curvature we now turn to the important case where it is constant:

Definition 4.6 *A connected semi-Riemannian manifold (M,g) has* **constant curvature** *if all sectional curvatures $C_x(X,Y)$ coincide (where $x \in M$ and $X,Y \in T_xM$ vary).*

We assume $n := \dim(M) \geq 2$. If $n \geq 3$, and $C_x(X,Y)$ is independent of X and Y for each x, then the common value of $C_x(X,Y)$ is also independent of x, so that (M,g) has constant curvature.[158]

Proposition 4.7 *If (M,g) has constant curvature, then the Riemann tensor (4.14), the Ricci tensor (3.14), and the Ricci scalar (3.15)–see also §4.5–are given by, respectively,*

$$R_{ijkl} = k(g_{ik}g_{jl} - g_{il}g_{jk}); \qquad R_{ij} = (n-1)kg_{ij}; \qquad R = n(n-1)k. \tag{4.85}$$

where k is the common value of all sectional curvatures, called the **curvature** *of (M,g).*

Proof. Let $C_x(X,Y) = k(x)$ for all $X,Y \in T_xM$ and some $k \in C^\infty(M)$. In terms of the tensor

$$S_x(V,W,X,Y) = g_x(V,X)g_x(W,Y) - g_x(V,Y)g_x(W,X) \tag{4.86}$$

of type $(4,0)$, eq. (4.47) gives $\mathrm{Riem}_x(X,Y,X,Y) = k(x)S(X,Y,X,Y)$. But since the Riemann tensor is completely defined by its sectional curvatures, this implies $\mathrm{Riem}_x = k(x)S$. □

In $n = 2$ this just means that the scalar curvature is constant. Definition 4.6 becomes increasingly stringent in higher dimension, as T_xM contains an increasing number of plane whose sectional curvatures has to be constant, but this is balanced by the larger variety of possible manifolds and metrics, so that the classification is the same for any dimension $n \geq 2$. Even the Riemannian and the Lorentzian cases look strikingly similar, as we shall see. We start with the former.[159]

Theorem 4.8 *If $n \geq 2$, any (geodesically) complete and simply connected Riemannian manifold (M,g) with constant curvature k is isometrically isomorphic to one of the following spaces:*

- *$k = 1/\rho^2 > 0$: The n-dimensional* **sphere** *S^n_ρ with radius $\rho > 0$ in $\mathbb{R}^{n+1}$, i.e.,*

$$S^n_\rho := \left\{ (x_1,\ldots,x_{n+1}) \in \mathbb{R}^{n+1} \mid \sum_{i=1}^{n+1} x_i^2 = \rho^2 \right\}, \tag{4.87}$$

  with metric inherited from $\mathbb{R}^{n+1}$ with Euclidean *metric $\delta(X,Y) = \sum_{i=1}^{n+1} X^iY^i$.*

- *$k = 0$: The n-dimensional* **Euclidean space** *$\mathbb{R}^n$ with metric $\delta(X,Y) = \sum_{i=1}^{n} X^iY^i$.*

- *$k = -1/\rho^2 < 0$: The n-dimensional* **hyperboloid** *H^n_ρ in $\mathbb{R}^{n+1}$ with label $\rho > 0$ defined by*

$$H^n_\rho := \left\{ (x_0,x_1,\ldots,x_n) \in \mathbb{R}^{n+1} \mid -x_0^2 + \sum_{i=1}^{n} x_i^2 = -\rho^2, x_0 > 0 \right\}, \tag{4.88}$$

  with metric inherited from $\mathbb{R}^{n+1}$ with Minkowski *metric $\eta(X,Y) = -X^0Y^0 + \sum_{i=1}^{n} X^iY^i$.*

In both S^n_ρ and H^n_ρ, the geodesics are the intersections with a plane in $\mathbb{R}^{n+1}$ through zero.

[158] See footnote 660 in §A.5 for a proof, or Corollary 2.2.7 in Wolf (2011), which is a classic on spaces of constant curvature in any signature. For the Riemannian case see also the beautiful treatment by Vinberg *et al.* (1993).

[159] For us, saying that M is *simply connected* also implies, by convention, that M is *connected*.

In $n = 2$, where these spaces were first discovered,[160] H^2_ρ is often realized as the *Poincaré disc*,

$$D^2_\rho = \{(u,v) \in \mathbb{R}^2 \mid u^2 + v^2 < 4\rho^2\}, \tag{4.89}$$

equipped with the metric (which is already given by Riemann in his 1854 Habilitation lecture)

$$ds^2 = 4 \cdot \frac{du^2 + dv^2}{(1 + \frac{1}{4}k(u^2+v^2))^2}, \tag{4.90}$$

where $k = -1/\rho^2$. The Poincaré disc has been turned into art in a famous woodcut by Escher:[161]

Circle Limit IV (Heaven and Hell) by M.C. Escher, showing the Poincaré disc

Independently of Bolyai and Lobachevskii, Riemann probably found the hyperbolic metric as follows: stereographic projection of S^2_ρ from the north pole onto the $z = 0$ plane in $\mathbb{R}^3$, i.e.

$$(x,y,z) \mapsto (u,v,0); \qquad u = \frac{\rho x}{\rho - z}; \qquad v = \frac{\rho y}{\rho - z}, \tag{4.91}$$

where $(x,y,z) \neq (0,0,\rho)$, gives the same metric (4.90), but this time with $k = 1/\rho^2$. However, the later model (4.88) has the advantage that (for $n = 2$) its geodesics are simply the intersections of H^2_ρ with planes in $\mathbb{R}^3$ through the origin, exactly as for S^2_ρ (giving the great circles).

[160] As mentioned in the historical introduction, the $2d$ hyperbolic spaces were independently discovered by Bolyai and Lobachevskii in the 1830s and caused a revolution, in that Euclidean geometry no longer provided an absolute source of truth in mathematics, so that eventually the link between mathematics and reality came to be dropped. For $k < 0$, the need for something like an embedding in Minkowski space arises because Hilbert (1901) proved that it is impossible to isometrically embed D^2_ρ with its hyperbolic metric in $\mathbb{R}^3$, equipped with its usual (Euclidean) metric.

[161] Copyright: *The M.C. Escher Company, Baarn*. See also Wieting (2010) and footnote 486.

In order to discuss the Lorentzian counterpart of Theorem 4.8 we need the ***Lorentzian cover*** of a non-simply connected Lorentzian manifold (M,g): this is the universal cover $\tilde{M}$ of M equipped with the pullback metric $\tilde{g}=\pi^* g$ of the covering projection $\pi:\tilde{M}\to M$. This complication was not necessary in Theorem 4.8, since both $S^n_\rho \cong S^n$ and $H^n_\rho \cong \mathbb{R}^n$ are simply connected for $n\geq 2$.

Theorem 4.9 *If $n\geq 2$, any (geodesically) complete and simply connected Lorentzian manifold (M,g) with constant curvature k is isometrically isomorphic to one of the following spaces:*[162]

- $k=1/\rho^2>0$*: For $n>2$, the n-dimensional* **de Sitter space** dS^n_ρ *with $\rho>0$ in $\mathbb{R}^{n+1}$, i.e.*

$$dS^n_\rho := \left\{ (x_0,x_1,\ldots,x_n)\in\mathbb{R}^{n+1} \mid -x_0^2+\sum_{i=1}^n x_i^2=\rho^2 \right\}, \tag{4.92}$$

  with metric inherited from $\mathbb{R}^{n+1}$ with Minkowski *metric $\eta(X,Y)=-X^0Y^0+\sum_{i=1}^n X^iY^i$. For $n=2$, however, one should use the Lorentzian cover $\widetilde{S^2_\rho}$ (with associated metric).*

- $k=0$*: The n-dimensional* **Minkowski space-time** $(\mathbb{M}_n,\eta_n)$*, i.e. $\mathbb{R}^n$ with metric (3.3).*

- $k=-1/\rho^2<0$*: The Lorentzian cover $\widetilde{AdS^n_\rho}$ of* **anti de Sitter space** *with $\rho>0$, i.e.*

$$AdS^n_\rho := \left\{ (x_{-1},x_0,x_1,\ldots,x_{n-1})\in\mathbb{R}^{n+1} \mid -x_{-1}^2-x_0^2+\sum_{i=1}^{n-1} x_i^2=-\rho^2 \right\}, \tag{4.93}$$

  with metric inherited from $\mathbb{R}^{n+1}$ with metric $\chi(X,Y)=-X_{-1}Y_{-1}-X^0Y^0+\sum_{i=1}^{n-1}X^iY^i$.

In both dS^n_ρ and AdS^n_ρ, the geodesics are the intersections with a plane in $\mathbb{R}^{n+1}$ through zero.[163]

The proof of Theorems 4.8 and 4.9 is very long,[164] but below we sketch the main argument. This uses some Lie group theory. We first comment on the similarity between Theorems 4.8 and 4.9.

[162] De Sitter space was introduced by de Sitter (1917ab), and independently, along with anti de Sitter space, by Levi-Civita (1917b). De Sitter's papers were a response to Einstein (1917b), the paper that launched relativistic cosmology (and, one could say, theoretical cosmology altogether). This, in turn, arose from earlier conversations and correspondence between Einstein and Willem de Sitter (1872–1934), who was a prominent Dutch astronomer (based at Leiden) and also an accomplished mathematician. Einstein (1917b) is also (in)famous because Einstein introduced his cosmological constant λ in it. In the early parts of the paper he paves the way for λ by claiming (incorrectly) that it solves the well-known paradox in Newtonian cosmology that if matter is distributed uniformly in an infinite universe, the gravitational force at each point is infinite. But his real purpose was to rescue Mach's principle in the context of his new theory (see footnote 24). De Sitter invented his model (which is a solution to Einstein's vacuum field equations with positive cosmological constant) to (successfully) challenge this, which led to an interesting and historical significant debate (Smeenk, 2014). Because of the discovery, at the very end of the twentieth century, of an accelerated expansion of the universe (Kirshner, 2002), which requires $\lambda>0$ (now reinterpreted as "dark energy" and usually called Λ, as in the "ΛCDM Standard Model of Cosmology") it is now widely believed that we approximately live in a de Sitter universe. See also Kragh (2007) and Nussbaumer & Bieri (2009). The popularity of anti de Sitter space has also exploded after the discovery of the AdS/CFT correspondence. Useful references on (anti) de Sitter space range from Hawking & Ellis (1973), §5.2, to Moschella (2005).

[163] Instead of the ones in Theorem 4.9, also here one has disc-like realizations of these spaces, which are simply obtained by replacing the Euclidean metric du^2+dv^2 in (4.90) with the Minkowski metric du^2-dv^2, and similarly in higher dimension. However, the realizations of Theorem 4.9 are more widely used in GR.

[164] Both theorems are a special case of Theorems 2.4.4 and 2.4.9 in Wolf (2011). Their common generalization is as follows. For $s=0$, equip $\mathbb{R}^{n+1}$ with the Euclidean metric $g_0^{(n+1)}=\delta$. For $1\leq s<n$, take the indefinite metric $g_s^{(n+1)}(X,Y)=-\sum_{i=1}^s X^iY^i+\sum_{j=s+1}^{n+1}X^iY^i$. For $\rho>0$, define $S^{n,s}_\rho\subset\mathbb{R}^{n+1}$ as a quadric $-\sum_{i=1}^s x_i^2+\sum_{j=s+1}^{n+1}x_i^2=\rho^2$,

Namely, the assumption of (geodesic) completeness is much more natural in the Riemannian setting than it is in Lorentzian geometry, where it is violated in many realistic space-times (see chapter 6). Instead, a more natural completeness assumption would be *global hyperbolicity*, cf. §5.7. In fact, de Sitter space is globally hyperbolic but anti de Sitter space is not. See §5.10.

The need for the topological covering construction comes from the diffeomorphisms

$$dS^n_\rho \cong \mathbb{R}\times S^{n-1}; \tag{4.94}$$

$$(x_0,x_1,\dots,x_n)\mapsto\left(x_0,\frac{x_1}{\sqrt{\rho^2+x_0^2}},\dots,\frac{x_n}{\sqrt{\rho^2+x_0^2}}\right); \tag{4.95}$$

$$AdS^n_\rho \cong S^1\times\mathbb{R}^{n-1}; \tag{4.96}$$

$$(x_{-1},x_0,x_1,\dots,x_{n-1})\mapsto\left(\frac{x_{-1}}{\sqrt{\rho^2+\sum_{i=1}^{n-1}x_i^2}},\frac{x_0}{\sqrt{\rho^2+\sum_{i=1}^{n-1}x_i^2}},x_1,\dots,x_{n-1}\right). \tag{4.97}$$

so that $dS^2_\rho\cong\mathbb{R}\times S^1$ and hence $\widetilde{S^2_\rho}\cong\mathbb{R}^2$, and, for any $n\geq 2$, we obtain $\widetilde{AdS^n_\rho}\cong\mathbb{R}^n$.

Given Theorems 4.8 and 4.9, any other complete Riemannian or Lorentzian manifold with constant curvature can be constructed from the above spaces by forming quotients of M by discrete subgroups Γ of the isometry group of (M,g) that act freely and properly discontinuously on M.[165] In particular, the $2d$ de Sitter space dS^2_ρ has constant curvature $k=1/\rho^2$, and for any $n\geq 2$ the multiply connected anti de Sitter spaces AdS^n_ρ all have constant curvature $k=-1/\rho^2$.

Finally, for those familiar with Lie groups, we reformulate Theorems 4.8 and 4.9 in those terms. Spaces of constant curvature (and many other interesting Riemannian or Lorentzian manifolds with less symmetry) can be realized as ***homogeneous spaces*** (or ***coset spaces***). See Appendix A; we will restrict the discussion here to the points of direct interest.

An ***isometry*** of a metric g on M is a diffeomorphism φ of M such that

$$\varphi^*g=g \qquad\Leftrightarrow\qquad g_{\varphi(x)}(\varphi'_x(X),\varphi'_x(Y))=g_x(X,Y)\ \ \forall x\in M, X,Y\in T_xM. \tag{4.98}$$

The set of all such diffeomorphisms φ is the ***isometry group*** of (M,g), denoted by $\mathrm{Iso}(M,g)$. This is by definition a subgroup of the "infinite-dimensional" group $\mathrm{Diff}(M)$, but it can be shown that $\mathrm{Iso}(M,g)$ is a *finite-dimensional* Lie group in the compact-open topology.[166]

Let G be some subgroup of $\mathrm{Iso}(M,g)$ and suppose that G acts ***transitively*** on M (i.e. for each $x,y\in M$ there is $\gamma\in G$ such that $y=\gamma x$). Choosing some fixed $x'\in M$ with stabilizer

$$H=\{\gamma\in G\mid \gamma x'=x'\}. \tag{4.99}$$

with the metric induced from $g^{(n+1)}_s$, and let $H^{n,s}_\rho\subset\mathbb{R}^{n+1}$ be the quadric $-\sum_{i=1}^{s+1}x_i^2+\sum_{j=s+2}^{n+1}x_i^2=-\rho^2$, with the metric induced from $g^{(n+1)}_{s+1}$. Then $S^{n,s}_\rho$ and $H^{n,s}_\rho$ are semi-Riemannian manifolds of signature $(s,n-s)$ with constant curvatures $k=1/\rho^2$ and $k=-1/\rho^2$, respectively. In particular, $dS^n_\rho=S^{n,1}_\rho$ and $AdS^n_\rho=H^{n,1}_\rho$. Complete this list with the $k=0$ case in signature $(s,n-s)$, which is obviously $\mathbb{R}^n$ with metric $g^{(n)}_s$; for $s=1$ this is Minkowski space-time. Passing to the universal semi-Riemannian cover for $S^{n,s}_\rho$ if $s=n-1$ and for $H^{n,s}_\rho$ if $s=1$, up to isometry these are all complete simply connected semi-Riemannian manifolds of signature $(s,n-s)$ with constant curvature. In these realizations, the geodesics are once again the intersections with a plane in $\mathbb{R}^{n+1}$ through zero.

[165] We say that Γ acts ***freely*** on M if $\gamma x=x$ implies $x=e$, and ***properly discontinuously*** if each $x\in M$ has a nbhd U such that the set $\{\gamma\in\Gamma\mid\gamma(U)\cap U\neq\emptyset\}$ is finite; in particular, Γ-orbits cannot have any accumulation point. Wolf (2011) contains a complete solution of this problem, which is already very substantial for the hyperbolic space H^2_ρ.

[166] See e.g. O'Neill (1983), Theorem 9.32. The ***compact-open topology*** on a space of maps $F:X\to Y$ is generated by open sets of the form $C_{K,U}=\{F\mid F(K)\subset U\}$, where K is compact in X and U is open in Y.

This is a closed and hence Lie subgroup of G, and we obtain a diffeomorphism

$$\psi : M \xrightarrow{\cong} G/H; \qquad \psi(\gamma' x') = \gamma' H. \tag{4.100}$$

If we define the canonical action of G on G/H by $\gamma(\gamma' H) = (\gamma\gamma')H$, then ψ is ***equivariant***, in that $\psi(\gamma x) = \gamma\psi(x)$. We then equip G/H with the unique metric $g' = (\psi^{-1})^* g$ that makes ψ an isometry, namely $g = \psi^* g'$. The above G-action on G/H is then transitive and isometric.

Conversely, we start with a Lie group G and a closed subgroup $H \subset G$ and study possibly G-invariant metrics on G/H. This is done in See Appendix A, with the following conclusion:

Proposition 4.10 *1. There is a bijective correspondence between G-invariant metrics on G/H and $\mathrm{Ad}'(H)$-invariant metrics on $\mathfrak{g}/\mathfrak{h}$, and hence, if $\mathfrak{g} = \mathfrak{h} \oplus \mathfrak{p}$ (see below), on $\mathfrak{p}$.*

2. There is a unique G-invariant metric on G/H (up to scaling by a positive constant) iff the $\mathrm{Ad}'(H)$-action on $\mathfrak{g}/\mathfrak{h}$ (or, if applicable, on $\mathfrak{p}$) is irreducible.

Here $\mathfrak{g}$ and $\mathfrak{h}$ are the Lie algebras of G and H, respectively, with $\mathfrak{h} \subset \mathfrak{g}$. Any group G acts on itself by the ***adjoint action*** $\mathrm{Ad}_\gamma(\delta) = \gamma\delta\gamma^{-1}$. If G is a Lie group, this action defines a representation Ad' of G on its Lie algebra $\mathfrak{g}$, defined by $\mathrm{Ad}'_\gamma(X) = \gamma X \gamma^{-1}$ (this notation is justified since in Appendix A we define Lie groups and their Lie algebras as matrices). This action may, of course, be restricted to $H \subset G$, and it is easy to see that this restriction quotients to $\mathfrak{g}/\mathfrak{h}$. In our application to spaces with constant curvature, the vector space $\mathfrak{g}$ has a canonical decomposition

$$\mathfrak{g} = \mathfrak{h} \oplus \mathfrak{p}, \tag{4.101}$$

where (trivially) not only $\mathfrak{h}$, but also $\mathfrak{p}$ is invariant under Ad'_h for any $h \in H$ (if H is connected, this invariance requirement is equivalent to $[\mathfrak{h}, \mathfrak{p}] \subset \mathfrak{p}$). This is meant in Proposition 4.10.1.

The proof of Proposition 4.10 is based on two ideas. First, a G-invariant metric g' on a homogeneous space G/H is determined by its value at any given point of G/H, for which one takes H (seen as the equivalence class of $e \in G$ in G/H). Second, one has an isomorphism

$$T_H(G/H) \cong \mathfrak{g}/\mathfrak{h}, \tag{4.102}$$

which is both linear and H-equivariant, in that the linear H-action on $T_H(G/H)$ coming from the G-action on G/H is mapped to the H-action on $\mathfrak{g}/\mathfrak{h}$ mentioned above.[167] Thus any G-invariant metric g' on G/H is determined by its value g'_H at $H \in G/H$, i.e. by a metric on the tangent space (4.102). This metric is still constrained by G-invariance, whose "infinitesimal shadow" at H is the adjoint H-action Ad'_h on $\mathfrak{g}/\mathfrak{h}$. If this shadow is sufficiently large, g'_H is even determined by Ad'_h-invariance (up to a constant scale factor). In words, g' is both ***homogeneous***, i.e. "the same" if one *moves* from point to point by the G-action, and, if the second part of Proposition 4.10 applies, also ***isotropic*** in being "the same" in all directions from a *given* point of view. This uniqueness applies in particular to spaces of constant curvature, to which we now return.

Let $O(k,l) \subset GL(k+l,\mathbb{R})$ be the isometry group of the metric $g = \mathrm{diag}(-\overset{k}{\cdots}-,+\overset{l}{\cdots}+)$ on $\mathbb{R}^{k+l}$; elements of $O(k,l)$ are matrices $\gamma \in GL(k+l,\mathbb{R})$ that satisfy $\gamma^T g \gamma = g$. We will be interested in $k = 0$, $k = 1$, and $k = 2$ and write $O(l)$ for $O(0,l)$. Then the following holds:

[167] Let $L_h(\gamma H) = (h\gamma)H$ be the restriction of the G-action on G/H to $h \in H$. Then the pushforward L'_h maps $T_{\gamma H}(G/H)$ to $T_{(h\gamma)H}(G/H)$, and so for $\gamma = e$ we obtain a linear map $L_h : T_H(G/H) \to T_H(G/H)$.

- $O(n+1)$ acts transitively on S^n_ρ. The stabilizer of $(0,\dots,0,\rho) \in S^n_\rho$ is $O(n)$, seen as a subgroup of $O(n+1)$ under the obvious "upper left corner" embedding. Hence

$$S^n_\rho \cong O(n+1)/O(n), \tag{4.103}$$

  first as manifolds. But if $O(n+1)/O(n)$ is equipped with ρ^2 times its canonical $O(n+1)$-invariant metric (see below), then this diffeomorphism is promoted to an isometry.

- $O(1,n)$ acts transitively on H^n_ρ. This time, take $(\rho,0\dots,0)$, whose stabilizer is again $O(n)$, embedded in the "lower right corner" of $O(1,n)$. Thus, with similar comments,

$$H^n_\rho \cong O(1,n)/O(n). \tag{4.104}$$

- $O(1,n)$ also acts transitively on de Sitter space dS^n_ρ. Returning to $(0,\dots,0,\rho)$, we obtain

$$dS^n_\rho \cong O(1,n)/O(1,n-1). \tag{4.105}$$

- $O(2,n-1)$ acts transitively on anti de Sitter space AdS^n_ρ. Taking $(\rho,0\dots,0)$ yields

$$AdS^n_\rho \cong O(2,n-1)/O(1,n-1). \tag{4.106}$$

- Finally, for the flat Euclidean and Minkowski spaces we have

$$\mathbb{R}^n \cong E(n)/O(n); \qquad \textbf{\textit{(Euclidean)}} \tag{4.107}$$
$$\mathbb{R}^n \cong P(n)/O(1,n-1); \qquad \textbf{\textit{(Minkowski)}}, \tag{4.108}$$

  where the semidirect product $E(n) = O(n) \ltimes \mathbb{R}^n$ is the ***Euclidean group*** in dimension n, and likewise $P(n) = O(1,n-1) \ltimes \mathbb{R}^n$ is the ***Poincaré group*** in dimension n. These are the isometry groups of the Euclidean metric and the Minkowski metric on $\mathbb{R}^n$, respectively.

In the Riemannian case the denominator is always $H = O(n)$, whereas in the Lorentzian case it is the (Lorentz!) group $H = O(1,n-1)$. It turns out that case 2 of Proposition 4.10 applies: in fact, in both cases we have $\mathfrak{g}/\mathfrak{h} \cong \mathbb{R}^n$ and under this isomorphism the adjoint H-action is simply given by the defining action of H on $\mathbb{R}^n$ (which is certainly irreducible). Thus G-invariant metrics on all of the above spaces G/H are unique (up to scaling), and hence "canonical".

Corollary 4.11 *If $n = \dim(M) \geq 2$, then the following list (where each space G/H is equipped with its canonical G-invariant metric) gives all complete and simply connected spaces M of constant curvature, up to isometry and up to rescaling of the metric by a positive constant:*

- **Riemannian** *i) $k > 0$: $O(n+1)/O(n)$. ii) $k = 0$: $E(n)/O(n)$. iii) $k < 0$: $O(1,n)/O(n)$.*
- **Lorentzian** *i) $k > 0$: $O(1,n)/O(1,n-1)$ (for $n = 2$ one needs its Lorentzian cover). ii) $k = 0$: $P(n)/O(1,n-1)$. iii) $k < 0$: the Lorentzian cover of $O(2,n-1)/O(1,n-1)$.*

Finally, realizations of homogeneous spaces as G/H are not unique; one may have $G'/H' \cong G/H$ (think of $(G\times G)/G \cong G/\{e\}$). As a case in point, all of the above groups are disconnected: $O(l)$ has two (connected) components, of which $SO(l)$ is the one containing the identity, and $O(1,l)$ and $O(2,l)$ even have four. The above way of writing down the isomorphisms has the advantage that $O(n+1)$ is the full isometry group of S^n_ρ, and likewise $O(1,n)$ for both H^n_ρ and dS^n_ρ, and $O(2,n-1)$ for AdS^n_ρ. However, each isomorphism is also true if both groups in the quotient are replaced by their identity components, e.g. $S^n_\rho \cong SO(n+1)/SO(n)$, etc.

4.5 Ricci tensor and Ricci scalar

The Riemann tensor contains all information about curvature. There also exist weaker measures of curvature. The main actor in GR is the ***Ricci tensor***, which like the metric has type $(2,0)$:

$$\mathrm{Ric}(X,Y) := \sum_{a=1}^{n} \mathrm{Riem}(e_a,X,e_a,Y); \qquad \mathrm{Ric}_{ij} \equiv R_{ij} = R^{l}_{\ ilj} = g^{kl}R_{kilj}, \tag{4.109}$$

where (e_a) is any orthonormal frame.[168] Note that Ric is symmetric by (4.36). From Ric, we define the ***scalar curvature*** by

$$R := \sum_{a=1}^{n} \mathrm{Ric}(e_a,e_a) = \sum_{a,b=1}^{n} C(e_a,e_b) = g^{ij}R_{ij}, \tag{4.110}$$

where of course in the second sum the terms $a \neq b$ do not contribute and hence due to symmetry the sum just has $(n^2-n)/2$ terms. For example, in $n=3$ the Ricci scalar (at a point x) is the average of the sectional curvatures of the x-y, x-z, and y-z planes (within the tangent space T_xM).

Furthermore, the Ricci tensor defines two ***Einstein tensors***, most easily by their components

$$G_{ij} := R_{ij} - \tfrac{1}{2}g_{ij}R; \tag{4.111}$$

$$E_{ij} := R_{ij} - \frac{1}{n}g_{ij}R. \tag{4.112}$$

Physicists use G_{ij} because, as will be explained later, it emerges from the calculus of variations applied to the functional $g \mapsto \int_M R(g)$. Mathematicians, on the other hand, use E_{ij} because it is simply the traceless part of Ric (note that $g^{ij}E_{ij}=0$). Moreover, to explain the name, suppose

$$\mathrm{Ric} = \lambda g; \qquad R_{ij} = \lambda g_{ij}, \tag{4.113}$$

for some constant $\lambda \in \mathbb{R}$, in which case we say that (M,g) is an ***Einstein manifold***, and that g is an ***Einstein metric***. Then $R = \lambda \cdot n$ is constant and $\lambda = R/n$, so that (4.113) implies $E_{ij}=0$. In $d>2$, also the converse is true;[169] this follows from the Bianchi identity (4.25). Thus:

Proposition 4.12 *For $n>2$, a metric satisfies (4.113) iff its Einstein tensor (4.112) vanishes.*

The symmetries (4.36) enable one to count the number of independent components of the Riemann tensor in various dimensions n, namely $n^2(n^2-1)/12$ (check!). Therefore:

1. For $n=2$ the Riemann tensor has just one independent component R_{1212}, and also

$$g^{-1} = \begin{pmatrix} g^{11} & g^{12} \\ g^{21} & g^{22} \end{pmatrix} = \frac{1}{\det(g)} \begin{pmatrix} g_{22} & -g_{12} \\ -g_{12} & g_{11} \end{pmatrix}, \tag{4.114}$$

so that the Ricci tensor $R_{ij} = g^{kl}R_{kilj}$ must equal $R_{ij} = g_{ij}R_{1212}/\det(g)$. This gives

$$R_{ijkl} = \tfrac{1}{2}R(g_{ik}g_{jl} - g_{il}g_{jk}); \qquad R_{ij} = \tfrac{1}{2}g_{ij}R, \tag{4.115}$$

cf. (4.85). Hence $R_{1212} = \frac{1}{2}\det(g)\cdot R = \det(g)\cdot K$, where the Gaussian curvature K is given, either as a definition or as a theorem,[170] by one of the equivalent expressions

$$K = C(\partial_1,\partial_2) = R_{1212}/\det(g) = \tfrac{1}{2}R. \tag{4.116}$$

[168] Authors use various sign conventions for the Riemann tensor, but all Ricci tensors and scalars coincide.
[169] We will shortly see that $E_{ij}=0$ in $d=2$, where we know since Gauss that non-constant R is certainly possible.
[170] See §4.3, as well as e.g. Heckman (2017), Theorem 3.15.

2. For $n = 3$, the Riemann tensor has 6 independent components, as does the Ricci tensor! So these two must carry the same information.[171] This can be understood from linear algebra, as follows. If V has an inner product, any symmetric bilinear map $T : V \otimes V \to \mathbb{R}$ is equivalent to a self-adjoint linear map $\tilde{T} : V \to V$ via $T(v \otimes w) = \langle v, \tilde{T} w \rangle$. In particular, the Ricci tensor $\mathrm{Ric}_x : T_xM \otimes T_xM \to \mathbb{R}$ at a point $x \in M$ is equivalent to a linear map

$$\widetilde{\mathrm{Ric}_x} : T_xM \to T_xM; \qquad g_x(X, \widetilde{\mathrm{Ric}_x} Y) = \mathrm{Ric}_x(X,Y). \tag{4.117}$$

If $\dim(V) = 3$, then $\dim(\Lambda^3 V) = 1$, and any nonzero $\mathrm{Vol} \in \Lambda^3 V$ gives an isomorphism

$$\Lambda^2 V \stackrel{\cong}{\to} V^*; \qquad A \mapsto \hat{A}; \qquad A \wedge v = \hat{A}(v)\mathrm{Vol}. \tag{4.118}$$

This follows from a dimension count: if V has a basis (e_1, e_2, e_3), then $\Lambda^2 V$ has a basis $(e_1 \wedge e_2, e_2 \wedge e_3, e_3 \wedge e_1)$. One may then take $\omega = e_1 \wedge e_2 \wedge e_3$, take the basis $(\omega^1, \omega^2, \omega^3)$ of V^* dual to the basis of V (i.e. $\omega^i(e_j) = \delta^i_j$), so that if $A = A^{ij} e_i \wedge e_j \in \Lambda^2 V$, we find

$$\hat{A}_i = \varepsilon_{ijk} A^{jk}. \tag{4.119}$$

Here ε_{ijk} is the totally antisymmetric (Levi-Civita) symbol, that is, $\hat{A}_1 = A^{23}$, $\hat{A}_2 = -A^{13}$, and $\hat{A}_3 = A^{12}$. Dually, $V \cong (\Lambda^2 V)^* = \Lambda^2 V^*$ under $v \mapsto \hat{v}$, where $\hat{v}_{ij} = \varepsilon_{ijk} v^k$. Consequently, in $n = 3$ (only!), one has $\Lambda^2 T_xM \cong T^*_xM \cong T_xM$. This isomorphism also makes linear maps $\Lambda^2 T_xM \to \Lambda^2 T_xM$ and $T_xM \to T_xM$ equivalent, so that the maps (4.40) and (4.117) are essentially the same. If the Ricci tensor as in (4.117) is diagonalized by an orthonormal basis (e_1, e_2, e_3) of T_xM with eigenvalues $(\lambda_1, \lambda_2, \lambda_3)$, then the Riemann tensor as in (4.40) is diagonal with respect to the basis $(e_1 \wedge e_2, e_2 \wedge e_3, e_3 \wedge e_1)$ of $\wedge^2 T_xM$, with eigenvalues $(\lambda_1 + \lambda_2 - \lambda_3, \lambda_2 + \lambda_3 - \lambda_1, \lambda_1 - \lambda_2 + \lambda_3)$. The Ricci scalar is given by

$$R_x = \lambda_1 + \lambda_2 + \lambda_3. \tag{4.120}$$

Though derived from the Ricci tensor, an interesting tensor in $n = 3$ is the ***Cotton tensor***

$$C_{ijk} := \nabla_k R_{ij} - \nabla_j R_{ik} + \tfrac{1}{4}(\tilde{g}_{ik} \nabla_j R - \tilde{g}_{ij} \nabla_k \tilde{R}). \tag{4.121}$$

Much as the Riemann tensor detects if a space(-time) is flat, cf. Theorem 4.1, the Cotton tensor detects ***conformal flatness***. First, C is invariant under rescalings $g \mapsto \Omega^2 g$ (i.e. ***conformal transformations***), where $\Omega \in C^\infty(M)$ is strictly positive. Since it vanishes for flat metrics, it then also vanishes if $g = \Omega^2 \delta$ (or $\Omega^2 \eta$). The converse can also be proved. Hence: *a 3d space or space-time is conformally flat iff its Cotton tensor vanishes.*[172]

3. For $n = 4$ (the case of interest to physics), the Riemann tensor has 20 independent components, whereas the Ricci tensor only has 10. The geometric information in the Riemann tensor that is not passed on to the Ricci tensor is contained in the ***Weyl tensor***

$$W_{klij} := R_{klij} + (g_{k[j} R_{i]l} + g_{l[i} R_{j]k}) + \tfrac{1}{3}(R \cdot g_{k[i} g_{j]l}), \tag{4.122}$$

where $[\cdots]$ *antisymmetrizes the enclosed indices* (e.g. $g_{k[j} R_{i]l} = g_{kj} R_{il} - R_{ki} g_{jl}$). The Weyl tensor ("Weyl") has the same symmetries as the Riemann tensor, cf. (4.36) - (4.38), so that Weyl also has 10 independent components, like Riem. Everything just said in $n = 3$ about the Cotton tensor is now valid in $n = 4$ for the Weyl tensor (which vanishes in $n = 3$).

[171] In $n = 3$ one has $R^l_{ijk} = g^l_j R_{ik} + g_{ik} R^l_j - g_{jk} R^l_i - g^l_i R_{jk} + \frac{1}{2} R(g^l_i g_{jk} - g^l_j g_{ik})$. In terms of the ***Kulkarni–Nomizu product*** $(P \odot Q)_{ijkl} := P_{il} Q_{jk} + P_{jk} Q_{il} - P_{ik} Q_{jl} - P_{jl} Q_{ik}$, this reads $\mathrm{Riem} = \frac{1}{4} R(g \odot g) + E \odot g$, cf. (4.112).

[172] The original reference is Cotton (1899), see also Eisenhart (1926), §28. A modern treatment is Garcia *et al.* (2004). There is no analogue of the Cotton or Weyl tensors in $n = 2$, since every $2d$ metric is conformally flat.

4.6 Submanifolds and hypersurfaces

As we saw in §4.3, differential geometry started with the study of two-dimensional submanifolds Σ of $\mathbb{R}^3$ (i.e. surfaces) by Gauss. In GR, a crucial role will be played by various submanifolds $\tilde{M}$ of a four-dimensional Lorentzian manifold M. One may define a submanifold $\tilde{M}$ of a manifold M in two equivalent ways: either as a subset $\tilde{M} \subset M$ of M with certain (good) properties, or as a manifold $\tilde{M}$ in its own right (a concept already defined, of course) plus an explicit map $F : \tilde{M} \to M$ with equivalent good properties. The former leads to the latter by considering the inclusion map $F : \tilde{M} \hookrightarrow M$, whereas the latter leads to the former by identifying $\tilde{M}$ with its image $F(\tilde{M}) \subset M$ (which may lead to some confusion!). We put $n := \dim(M)$ as usual.

Definition 4.13 *If $\tilde{M}$ is a manifold, a map $F : \tilde{M} \to M$ defines a* **submanifold** *$F(\tilde{M}) \subset M$ iff:*

1. *F is a homeomorphism onto its image $F(\tilde{M})$. In particular, F is injective.*[173]
2. *$F'_u : T_u\tilde{M} \to T_{F(u)}M$ is injective for all $u \in \tilde{M}$. Equivalently, the rank of F'_u equals $\dim(\tilde{M})$.*

Equivalently,[174] *a subset $\tilde{M} \subset M$ is a* ***k*-dimensional submanifold** *of M iff each $x \in \tilde{M}$ has an open nbhd U in M for which there is chart $\varphi : U \xrightarrow{\cong} V \subset \mathbb{R}^n$ (for M) whose image takes the form*

$$\varphi(U \cap \tilde{M}) = \varphi(U) \cap X, \tag{4.123}$$

where X is a k-dimensional affine linear subspace of $\mathbb{R}^n$. If $k = n-1$, $\tilde{M}$ is called a **hypersurface**.

Until the end of this chapter we assume $\tilde{M}$ is a hypersurface, i.e. $\dim(\tilde{M}) = n-1$. If M carries a metric tensor g, then, generalizing (4.50) in §4.3, $\tilde{M}$ inherits a–not necessarily metric!–tensor

$$\tilde{g} := \imath^* g; \qquad\qquad \tilde{g} \in \mathfrak{X}^{(2,0)}(\tilde{M}), \tag{4.124}$$

defined by the inclusion $\imath : \tilde{M} \hookrightarrow M$. Identifying $\tilde{M}$ with $\imath(\tilde{M})$, this simply means that

$$\tilde{g}_x(X_x, Y_x) = g_x(X_x, Y_x), \tag{4.125}$$

for any $X_x, Y_x \in T_x\tilde{M} \subset T_xM$, with $x \in \tilde{M}$. It is easy to see that if (M, g) is Riemannian, then so is $(\tilde{M}, \tilde{g})$. But in the Lorentzian case the induced "metric" $\tilde{g}$ need not be non-degenerate.

Lemma 4.14 *Let $g : V \times V \to \mathbb{R}$ be a symmetric nondegenerate bilinear map on a real vector space V. Then for any linear subspace $W \subset V$, with $W^\perp := \{v \in V \mid g(v, w) = 0 \forall w \in W\}$,*

$$\dim(W) + \dim(W^\perp) = \dim(V); \tag{4.126}$$

$$(W^\perp)^\perp = W. \tag{4.127}$$

For the proof see O'Neill (1983), Lemma 2.22. Taking $V = T_xM$ and $W = T_x\tilde{M}$, this yields

$$\dim((T_x\tilde{M})^\perp) = 1. \tag{4.128}$$

Hence at each $x \in \tilde{M}$ one has a *normal (vector)* $N_x \in (T_x\tilde{M})^\perp \subset T_xM$ i.e. $g(N_x, X_x) = 0$ for all $X_x \in T_x\tilde{M}$, which by (4.128) is unique up to scalar multiplication (but we assume $N_x \neq 0$).

If (M, g) is Riemannian, then we may normalize each N_x such that

$$g_x(N_x, N_x) = 1. \tag{4.129}$$

[173] Dropping this condition defines an ***immersed submanifold***; what we define is an ***embedded submanifold***.

[174] See e.g. Andrews (undated), Proposition 3.2.1, combined with Proposition 1.31 in O'Neill (1983).

This still yields no canonical choice of N_x, but any two choices only differ by a sign and we assume that we can make a smooth choice $x \mapsto N_x$ throughout $\tilde{M}$, which is always Riemannian.[175]

The Lorentzian case is much richer. A hypersurface may not fall into any of the three classes below since the sign of $g_x(N_x, N_x)$ may change with x, but let as assume it is fixed (or zero).

Definition 4.15 *A hypersurface $\tilde{M} \subset M$, with nonzero normal vector field N, is called:*

- **spacelike** *iff $g_x(N_x, N_x) < 0$ for each $x \in \tilde{M}$. This is the case iff the induced metric $g_{|\tilde{M}}$ is positive definite, so that $(\tilde{M}, g_{|\tilde{M}})$ is a 3d Riemannian manifold.*
- **timelike** *or* **Lorentzian** *iff $g_x(N_x, N_x) > 0$ for each $x \in \tilde{M}$. This is the case iff the induced metric $g_{|\tilde{M}}$ is Lorentzian (obviously with signature $(-++)$).*
- **null** *iff $g_x(N_x, N_x) = 0$ for each $x \in \tilde{M}$. This is the case iff $g_{|\tilde{M}}$ is degenerate.*

To explain the last remark, we first note that in the null case, N_x is both normal and tangent to its null hypersurface. To see this, take $W = T_x\tilde{M}$ in Lemma 4.14, so that $W^\perp = \mathbb{R} \cdot N_x$, whence

$$T_x\tilde{M} = (T_x\tilde{M}^\perp)^\perp = (\mathbb{R} \cdot N_x)^\perp. \tag{4.130}$$

Hence $N_x \in T_x\tilde{M}$, but by definition of a normal, there exists no $X_x \in T_x\tilde{M}$ for which $g(N_x, X_x) \neq 0$, so that $g_{|\tilde{M}}$ is degenerate. Furthermore, whereas *timelike* normals are usually normalized as

$$g_x(N_x, N_x) = -1, \tag{4.131}$$

and *spacelike* normals usually satisfy (4.129), in the *null* case the normals N_x lack a natural normalization. To soften this, note that $T_x\tilde{M}$ cannot contain any null vector N'_x linearly independent of N_x (for in that case $g_x(N_x, N'_x) = 0$, which would contradict the Lorentzian signature of g). Therefore, we can find a second null vector field $\underline{N}_x \in T_xM$ (pointing outside $T_x\tilde{M}$) such that

$$g_x(N_x, \underline{N}_x) = -1. \tag{4.132}$$

Lemma 4.16 *If $\tilde{M}$ is null, any $X_x \in T_x\tilde{M}$ is either proportional to N_x (hence null), or spacelike.*

Proof. Suppose $T_x\tilde{M}$ contains a timelike vector T_x; then $g(T_x + \lambda N_x) = g(T_x, T_x) < 0$ for all λ, but a computation in coordinates shows that any sum $T_x + \lambda N_x$ of a timelike vector and a null vector becomes spacelike for large λ. The claim follows by the argument after (4.131). □

This is important, because it shows that *null hypersurfaces have a canonical lightlike direction* given by its normal (!); see §5.3 and §6.3 for further discussion, especially Proposition 6.9.

There are two basic examples of *null* hypersurfaces in Minkowski space-time $(\mathbb{M}, \eta)$. On the one hand, we have null *hyperplanes* such as $u := t - r$ or $v := t + r$ constant, or more generally the set of all vectors orthogonal to a given null vector (and translates thereof). On the other hand, we have forward or backward *lightcones*, see §5.3. In GR, in the context of black holes event horizons and Cauchy horizons are null hypersurfaces (see §10.7), and null hypersurfaces also play an important role in the settting of Penrose's singularity theorem (see §6.3).

Spacelike hypersurfaces are also very important in GR, especially Cauchy surfaces; the simplest example in $\mathbb{M}$ is x^0 = constant. Similarly, for a *timelike* hypersurface we may take x^i = constant for $i = 1$, 2, or 3. A more spectacular example in GR is the photon sphere in Schwarzschild space-time (see §9.2), and also "naked singularities" are timelike (see chapter 9).

[175] A sufficient condition for this, i.e. triviality of the normal bundle, is that $\tilde{M}$ be connected and simply connected (Kobayashi & Nomizu, 1969, p. 5). Since the criteria in Definition 4.15 are independent of the sign of N_x, by Lemma 4.14 the classification in this definition is even well defined if no continuous choice $x \mapsto N_x$ exists on $\tilde{M}$.

4.7 Gauss–Weingarten and Gauss–Codazzi equations

Let $\tilde{M} \subset M$ be a hypersurface with normal N; if M is Lorentzian we assume $\tilde{M}$ is spacelike (this case is fundamental to GR). The orthogonal projection from T_xM onto $T_x\tilde{M}$ is given by

$$\pi_x : T_xM \to T_x\tilde{M} \subset T_xM; \tag{4.133}$$
$$\pi_x(X_x) = X_x - g_x(X_x,N_x)N_x; \quad \text{(Riemannian case)}; \tag{4.134}$$
$$\pi_x(X_x) = X_x + g_x(X_x,N_x)N_x; \quad \text{(Lorentzian spacelike case)}, \tag{4.135}$$

so that $\pi_x(N_x) = 0$ and $\pi_x(X_x) = X_x$ if $X_x \in T_x\tilde{M}$ (this projection is independent of the choice of N_x). Then (M,g) and $(\tilde{M},\tilde{g})$ each have their Levi-Civita connections ∇ and $\tilde{\nabla}$, respectively.

Proposition 4.17 *The connection ∇ on M is related to the connection $\tilde{\nabla}$ on $\tilde{M}$ by*

$$\pi(\nabla_X Y) = \tilde{\nabla}_X Y \quad (X,Y \in \mathfrak{X}(\tilde{M})). \tag{4.136}$$

Here the covariant derivative $\tilde{\nabla}_X Y$ on the right-hand side is clearly defined (as an element of $\mathfrak{X}(\tilde{M})$), but also the covariant derivative $\nabla_X Y$ in M on the left-hand side is well defined, even though Y is merely a vector field on $\tilde{M}$ rather than on all of M: as in the comment preceding (3.40), if $X \in \mathfrak{X}(\tilde{M})$ and $Y \in \mathfrak{X}(M)$, then the value of $\nabla_X Y$ only depends on the restriction of Y to $\tilde{M}$ (indeed, it only depends on the values of Y along the flow lines on X, which lie in $\tilde{M}$), and so $\nabla_X Y$ is defined (as a vector field on $\tilde{M}$) even when $Y \in \mathfrak{X}(\tilde{M})$.[176]

Proof. We write $\nabla'_X Y$ for $\pi(\nabla_X Y)$, so that (in the Lorentzian case for simplicity)

$$\nabla'_X Y = \nabla_X Y + g(\nabla_X Y, N)N. \tag{4.137}$$

We first check that ∇' is a covariant derivative on $\mathfrak{X}(\tilde{M})$. Linearity in Y is obvious (since both g and ∇_X are linear), as is $C^\infty(\tilde{M})$-linearity, cf. (3.32). The Leibniz rule (3.33) follows from the corresponding rule for ∇ and the property $g((Xf)Y,N) = (Xf)g(Y,N) = 0$ (since $Y \in \mathfrak{X}(\tilde{M})$). To identify ∇' with $\tilde{\nabla}$, we need to check that ∇' is torsion-free and metric. First,

$$\begin{aligned} \nabla'_X Y - \nabla'_Y X &= \nabla_X Y - \nabla_Y X + g(\nabla_X Y - \nabla_Y X, N)N \\ &= [X,Y] + g([X,Y],N)N = [X,Y], \end{aligned} \tag{4.138}$$

since ∇ (being the Levi-Civita connection on TM) is torsion-free, and $[X,Y] \in \mathfrak{X}(\tilde{M})$, assuming $X,Y \in \mathfrak{X}(\tilde{M})$, so that $g([X,Y],N) = 0$. Second, ∇' should satisfy (3.50), i.e.

$$X(\tilde{g}(Y,Z)) = \tilde{g}(\nabla'_X Y, Z) + \tilde{g}(Y, \nabla'_X Z) \qquad (X,Y,Z \in \mathfrak{X}(\tilde{M})). \tag{4.139}$$

This is quite obvious, since for $X,Y,Z \in \mathfrak{X}(\tilde{M})$ we have

$$\tilde{g}(\nabla'_X Y, Z) = g(\nabla'_X Y, Z) = g(\nabla_X Y + g(\nabla_X Y, N)N, Z) = g(\nabla_X Y, Z), \tag{4.140}$$

since $g(N,Z) = 0$, and so the right-hand side of (4.139) equals $g(\nabla_X Y, Z) + g(Y, \nabla_X Z)$. By (3.50) for ∇ and g, this in turn equals $X(g(Y,Z)) = X(\tilde{g}(Y,Z))$. This gives (4.139).

The claim now follows from Theorem 3.9. □

[176] In other words, if one insists that $\nabla_X : \mathfrak{X}(M) \to \mathfrak{X}(M)$, one may extend $Y \in \mathfrak{X}(M)$ to an arbitrary vector field on M, and if $X \in \mathfrak{X}(\tilde{M})$, then $\nabla_X Y$ is independent of this extension.

Eq. (4.136) implies the general ***Gauss–Weingarten equations***, where still $X,Y \in \mathfrak{X}(\tilde{M})$:

$$\nabla_X Y = \tilde{\nabla}_X Y + \tilde{k}(X,Y)N \quad \text{(Riemann)}; \tag{4.141}$$

$$\nabla_X Y = \tilde{\nabla}_X Y - \tilde{k}(X,Y)N \quad \text{(Lorentz)}; \tag{4.142}$$

$$\nabla_X N =: -W(X). \tag{4.143}$$

Here we take (4.143) to be the *definition* of the ***Weingarten map*** $W_x : T_x\tilde{M} \to T_x\tilde{M}$, noting that since $g(N,N) = \pm 1$, we have $g(\nabla_X N, N) = 0$ and hence $\nabla_X N \in T\tilde{M}$. Furthermore,

$$\tilde{k}(X,Y) := g(W(X),Y) = -g(\nabla_X N, Y) \tag{4.144}$$

defines the ***extrinsic curvature*** $\tilde{k} \in \mathfrak{X}^{(2,0)}(\tilde{M})$. As in (4.59), from the property

$$g(\nabla_X Y, N) = -g(Y, \nabla_X N), \tag{4.145}$$

which is proved as in the text between (4.58) and (4.59), we infer that $\tilde{k}$ is symmetric, viz.

$$\tilde{k}(X,Y) = -g(\nabla_X N, Y) = g(N, \nabla_X Y) = g(N, \nabla_Y X) = \tilde{k}(Y,X). \tag{4.146}$$

Eqs. (4.141) - (4.142) then easily follow from (4.136), giving the ("parallel") component in $T\tilde{M}$, and from taking the inner product with N, using (4.129) - (4.131), giving the normal component.

We also derive the general ***Gauss–Codazzi equations***, which, for $W,X,Y,Z \in \mathfrak{X}(\tilde{M})$, are:

$$\mathrm{Riem}(W,Z,X,Y) = \widetilde{\mathrm{Riem}}(W,Z,X,Y) + \tilde{k}(W,Y)\tilde{k}(X,Z) - \tilde{k}(W,X)\tilde{k}(Y,Z) \quad \text{(R)}; \tag{4.147}$$

$$\mathrm{Riem}(W,Z,X,Y) = \widetilde{\mathrm{Riem}}(W,Z,X,Y) + \tilde{k}(W,X)\tilde{k}(Y,Z) - \tilde{k}(W,Y)\tilde{k}(X,Z) \quad \text{(L)}; \tag{4.148}$$

$$\mathrm{Riem}(N,Z,X,Y) = (\tilde{\nabla}_X \tilde{k})(Y,Z) - (\tilde{\nabla}_Y \tilde{k})(X,Z), \tag{4.149}$$

where $\mathrm{Riem} \in \mathfrak{X}^{(3,1)}(M)$ and $\widetilde{\mathrm{Riem}} \in \mathfrak{X}^{(3,1)}(\tilde{M})$ are the Riemann curvature tensor for the Levi-Civita connection ∇ on TM (for g) and $\tilde{\nabla}$ on $T\tilde{M}$ (for $\tilde{g}$), respectively. The Codazzi relation (4.149) is the same for the Riemannian and the Lorentzian cases. These equations follow from two computations, which we perform for the Lorentzian case, i.e. using (4.142). The first is:

$$\begin{aligned}
\nabla_X \nabla_Y Z &= \nabla_X (\tilde{\nabla}_Y Z - \tilde{k}(Y,Z)N) \\
&= \tilde{\nabla}_X \tilde{\nabla}_Y Z - \tilde{k}(X, \tilde{\nabla}_Y Z)N - X(\tilde{k}(Y,Z)) \cdot N - \tilde{k}(Y,Z)\nabla_X N \\
&= \tilde{\nabla}_X \tilde{\nabla}_Y Z + W(X)\tilde{k}(Y,Z) - (\tilde{k}(X, \tilde{\nabla}_Y Z) + X(\tilde{k}(Y,Z)))N.
\end{aligned} \tag{4.150}$$

The second computation, which uses torsion-freeness of $\tilde{\nabla}$, i.e. $\tilde{\nabla}_X Y - \tilde{\nabla}_Y X = [X,Y]$, is

$$\nabla_{[X,Y]} Z = \tilde{\nabla}_{[X,Y]} Z - \tilde{k}([X,Y],Z)N = \tilde{\nabla}_{[X,Y]} Z - (\tilde{k}(\tilde{\nabla}_X Y, Z) - \tilde{k}(\tilde{\nabla}_Y X, Z))N. \tag{4.151}$$

The definition (4.10) of curvature, combined with the "covariant Leibniz rule"

$$X(\tilde{k}(Y,Z)) = (\tilde{\nabla}_X \tilde{k})(Y,Z) + \tilde{k}(\tilde{\nabla}_X Y, Z) + \tilde{k}(Y, \tilde{\nabla}_X Y), \tag{4.152}$$

which is a special case of (3.65), then yields, after some neat cancellations:[177]

$$\begin{aligned}
\Omega(X,Y)Z = (\nabla_X \nabla_Y - \nabla_Y \nabla_X - \nabla_{[X,Y]})Z = \tilde{\Omega}(X,Y)Z \\
+ W(X)\tilde{k}(Y,Z) - W(Y)\tilde{k}(X,Z) + ((\tilde{\nabla}_Y \tilde{k})(X,Z) - (\tilde{\nabla}_X \tilde{k})(Y,Z))N.
\end{aligned} \tag{4.153}$$

Taking the (metric) inner product with W and using (4.144) yields ***Gauss's equation*** (4.148), whereas the inner product with N and using (4.131) yields ***Codazzi's equation*** (4.149).

[177] Recall that unlike $\tilde{k}$, the metric is covariantly constant, i.e. $\tilde{\nabla}_X \tilde{g} = 0$ for all $X \in \mathfrak{X}(\tilde{M})$, cf. (3.67).

4.8 Fundamental theorem for hypersurfaces

The classical theory culminates in the ***fundamental theorem for hypersurfaces***, which was proved (by different means) in the 19th century. We discuss the proof in some detail,[178] since it will turn out to be a good preparation for the 3+1 split of the Einstein equations later on.

Theorem 4.18 *Let $(\tilde{M},\tilde{g})$ be a connected and simply connected m-dimensional Riemann manifold equipped with a symmetric tensor $\tilde{k}\in\mathfrak{X}^{(2,0)}(\tilde{M})$ satisfying the Gauss–Codazzi equations*

$$\widetilde{\mathrm{Riem}}(W,Z,X,Y)+\tilde{k}(W,Y)\tilde{k}(X,Z)-\tilde{k}(W,X)\tilde{k}(Y,Z)=0; \tag{4.154}$$

$$(\tilde{\nabla}_X\tilde{k})(Y,Z)-(\tilde{\nabla}_Y\tilde{k})(X,Z)=0. \tag{4.155}$$

Then there exists an isometric embedding $F:\tilde{M}\to\mathbb{R}^{m+1}$ for which the extrinsic curvature is the given tensor $\tilde{k}$. Such an embedding is unique up to isometry, which in the case at hand (i.e. $\mathbb{R}^{m+1}$ with Euclidean metric) means: up to combinations of translations, rotations, and reflections.

Note that (4.154) - (4.155) arise from (4.147) - (4.149) by putting Riem $=0$ (because $\mathbb{R}^{m+1}$ is equipped with the flat Euclidean metric), and have (4.79) - (4.80) as their coordinate version. The latter were admittedly written down and derived for $m=2$, but simply letting the indices α,β etc. run from 1 to m rather than from 1 tot 2 immediately generalizes our treatment of the classical theory of surfaces to any dimension (alas with some loss of visualisability).

We just prove a local version of Theorem 4.18 by PDE methods, which is enough to show the role of the Gauss–Codazzi equations as integrability conditions. So let us initially assume we found an $F:U\to\mathbb{R}^{m+1}$ satisfying the conditions in the theorem, where $U\in\tilde{M}$ is open. We make F unique by imposing the conjunction of the following local conditions:

1. For arbitrary $u_0\in U$ and $x_0\in\mathbb{R}^{m+1}$, the map F satisfies $F(u_0)=x_0$;
2. For some fixed orthonormal basis $(e_1,\ldots,e_m)$ of $T_{u_0}\tilde{M}$ and some given orthonormal basis $(f_1,\ldots,f_{m+1})$ of $T_{x_0}\mathbb{R}^{m+1}\cong\mathbb{R}^{m+1}$, its derivative satisfies $F'_{u_0}(e_\alpha)=f_\alpha$ $(\alpha=1,\ldots,m)$.

Without loss of generality we may choose geodesic normal coordinates on U relative to u_0, cf. (5.33) - (5.38) below, so that $e_\alpha=\partial_\alpha\equiv\partial/\partial u^\alpha$ is indeed orthonormal at least at u_0. Furthermore, we may pick coordinates (x^i) on $\mathbb{R}^{m+1}$ $(i=1,\ldots,m+1)$ such that $f_i=\partial/\partial x^i$ for $i=1,\ldots,m$. The components $F^i(u^\alpha)$ of $F:U\to\mathbb{R}^{m+1}$ then satisfy the (initial) condition

$$\frac{\partial F^i}{\partial u^\alpha}(u_0)=\delta^i_\alpha\ \ (\alpha=1,\ldots,m,\,i=1,\ldots,m); \tag{4.156}$$

$$\frac{\partial F^{m+1}}{\partial u^\alpha}(u_0)=0\ \ (\alpha=1,\ldots,m). \tag{4.157}$$

In addition to F, we have to define a normal vector field $\vec{N}$ on U, whose components N^i satsify

$$N^i(u_0)=0\ \ (i=1,\ldots,m); \tag{4.158}$$

$$N^{m+1}(u_0)=1. \tag{4.159}$$

[178] Cf. Kobayashi & Nomizu (1969), §VII.7, considerably rewritten. The argument uses some exterior calculus. For general $\tilde{M}$ the above theorem holds at least locally, in that any $u_0\in\tilde{M}$ has a connected and simply connected neighbourhood $U\in\mathscr{O}(\tilde{M})$ for which the above claims hold.

If we recall (3.61), whose asterisk we omit, for each $i = 1,\dots,m+1$ we have

$$(\tilde{\nabla}_\alpha dF^i)_\beta = x^i_{\alpha\beta} - \tilde{\Gamma}^\gamma_{\alpha\beta} x^i_\gamma, \tag{4.160}$$

where $\tilde{\nabla}_\alpha := \tilde{\nabla}_{\partial/\partial u^\alpha}$. Therefore, introducing 1-forms $\theta^i \in \Omega(U)$ for each $i = 1,\dots,m+1$ via

$$\theta^i = dF^i, \tag{4.161}$$

Gauss's equation (4.77) for $(\vec{x}_\alpha)$ turns (4.160) with (4.161) into

$$(\tilde{\nabla}_\alpha \theta^i)_\beta = \tilde{k}_{\alpha\beta} N^i \ \ (\alpha,\beta = 1,\dots,m). \tag{4.162}$$

Conversely, if $\theta^i \in \Omega(U)$ satisfies (4.162), there exists $F^i \in C^\infty(U)$ such that (4.161) holds. We start with a computation for any $\theta^i \in \Omega(U)$, which uses the Leibniz rule (3.65):[179]

$$\begin{aligned} d\theta^i(X,Y) &= X(\theta^i(Y)) - Y(\theta^i(X)) - \theta^i([X,Y]) \\ &= (\tilde{\nabla}_X \theta^i)(Y) + \theta^i(\tilde{\nabla}_X Y) - (\tilde{\nabla}_Y \theta^i)(X) - \theta^i(\tilde{\nabla}_Y X) - \theta^i([X,Y]) \\ &= (\tilde{\nabla}_X \theta^i)(Y) - (\tilde{\nabla}_Y \theta^i)(X) + \theta^i(\tau(X,Y)) \\ &= (\tilde{\nabla}_X \theta^i)(Y) - (\tilde{\nabla}_Y \theta^i)(X), \end{aligned} \tag{4.163}$$

since the Levi-Civita connection $\tilde{\nabla}$ is torsion-free, cf. (3.43). Eq. (4.162) then gives

$$d\theta^i(\partial_\alpha,\partial_\beta) = (\tilde{\nabla}_\alpha \theta^i)(\partial_\beta) - (\tilde{\nabla}_\beta \theta^i)(\partial_\alpha) = N^i(\tilde{k}_{\alpha\beta} - \tilde{k}_{\beta\alpha}) = 0, \tag{4.164}$$

by symmetry of the extrinsic curvature $\tilde{k}$. The Poincaré lemma then gives (4.161).

It is convenient to replace the 1-forms θ^i by the corresponding vector fields $Z^i = \sharp(\theta^i)$ on U ($i = 1,\dots,m+1$), in terms of which (4.162) becomes, writing Z^β_i for $(Z^i)^\beta$:

$$\frac{\partial Z^\beta_i}{\partial u^\alpha} + \tilde{\Gamma}^\beta_{\alpha\gamma} Z^\gamma_i = N^i \tilde{k}^\beta_\alpha. \tag{4.165}$$

Similarly, in terms of Z_i, Weingarten's equation (4.78) becomes

$$\frac{\partial N^i}{\partial u^\alpha} = -\tilde{k}_{\alpha\beta} Z^\beta_i. \tag{4.166}$$

We may rewrite the coupled PDEs (4.165) and (4.166) on U, $i = 1,\dots,m+1$, more elegantly as

$$\tilde{\nabla}_X Z^i = N^i W(X); \tag{4.167}$$

$$XN^i = -\tilde{k}(X,Z^i), \tag{4.168}$$

for $X \in \mathfrak{X}(U)$ and $N^i \in C^\infty(U)$, subject to the initial conditions (4.158) - (4.159) for N^i, with

$$Z^\alpha_i(u_0) = \delta^\alpha_i \ \ (\alpha = 1,\dots,m,\, i = 1,\dots,m); \tag{4.169}$$

$$Z^\alpha_{m+1}(u_0) = 0 \ \ (\alpha = 1,\dots,m). \tag{4.170}$$

We derived (4.167) - (4.168) with (4.169) - (4.170) from the existence of $F : U \to \mathbb{R}^{m+1}$ with the desired properties (as stated in the theorem). Conversely, if we can solve these equations for Z^i (and N^i), we are able to construct F, having the right properties, via $\theta^i = \flat(Z^i)$ and (4.161).

[179] In the first line we use the identity $d\omega(X,Y) = X(\omega(Y)) - Y(\omega(X)) - \omega([X,Y])$, valid for any $\omega \in \Omega(U)$.

We now show that this can be done. To begin with, we show that the integrability conditions for (4.167) - (4.168) are the Gauss–Codazzi equations, which should come as no surprise, since (4.167) - (4.168) are a version of the Gauss–Weingarten equations. From (4.168) we derive

$$[X,Y]N^i = -X\tilde{k}(Y,Z^i) + Y\tilde{k}(X,Z^i); \tag{4.171}$$
$$[X,Y]N^i = -\tilde{k}([X,Y],Z^i). \tag{4.172}$$

so that $X\tilde{k}(Y,Z^i) - Y\tilde{k}(X,Z^i) = \tilde{k}([X,Y],Z^i)$; a computation very similar to (4.163) then rewrites this as Codazzi's eq. (4.155). Similarly, practically the same computation as (4.150) - (4.153), using (4.155), shows that (4.167) implies Gauss's eq. (4.154). Thus the Gauss–Codazzi equations are *necessary* for the solvability of (4.167) - (4.168), which explains their role in Theorem 4.18.

To show that they are also *sufficient*, we have to make our hands dirty (as usual in PDE theory). We take geodesic normal coordinates (u^α) relative to $u_0 \in U$ (it may be necessary to shrink U in order to make it a normal nbhd) and some fixed orthonormal basis $(e_1,\dots,e_m)$ of $T_{u_0}\tilde{M}$, so that the coordinates $(u^1,\dots,u^m)$ specify the point $u = \gamma_{\vec{u}}(1)$, where $\gamma_{\vec{u}}$ is the (unique) geodesic having $\gamma_{\vec{u}}(0) = u_0$ and $\dot{\gamma}_{\vec{u}}(0) = u^\alpha e_\alpha$ (summation convention!).

For *fixed* $u \in U$, define a vector field Z^i and functions N^i along this geodesic $\gamma_{\vec{u}}$ by solving

$$\tilde{\nabla}_{\dot{\gamma}_{\vec{u}}} Z^i = N^i W(\dot{\gamma}_{\vec{u}}); \tag{4.173}$$
$$\dot{\gamma}_{\vec{u}} N^i = -\tilde{k}(\dot{\gamma}_{\vec{u}}, Z^i), \tag{4.174}$$

at least for $t \in [0,1]$, or, in coordinates, where $Z^i = (Z^1_i,\dots,Z^m_i)$ as above, and $tu = \gamma_{\vec{u}}(t)$,

$$\frac{dZ^\beta_i(t)}{dt} + u^\gamma \tilde{\Gamma}^\beta_{\gamma\alpha}(tu) Z^\alpha_i(t) = N^i(t)\tilde{k}^\beta_\alpha(tu)u^\alpha; \tag{4.175}$$
$$\frac{dN_i(t)}{dt} = -\tilde{k}_{\alpha\beta}(tu)u^\alpha Z^\beta_i, \tag{4.176}$$

with initial conditions $Z^\alpha_i(0) = \delta^\alpha_i$ $(i \le m)$, $Z^\alpha_{m+1}(0) = 0$, $N^i(0) = 0$ $(i \le m)$, and $N_{m+1}(0) = 1$, cf. (4.169) - (4.170) and (4.158) - (4.159). Here we identified $Z^i(t)$ with $Z^i(tu)$, etc. By standard ODE theory, $Z^i(t)$ and $N^i(t)$ exist and are unique. Finally, define $Z^i \in \mathfrak{X}(U)$ and $N^i \in C^\infty(U)$ by

$$Z^i(u) = Z^i(1); \tag{4.177}$$
$$N^i(u) = N^i(1), \tag{4.178}$$

where the Z^i and N^i on the right-hand side depend on u by construction. We claim that this pair (Z^i,N^i) solves (4.167) - (4.168) with the right initial conditions (4.169) - (4.170) and (4.158) - (4.159). To prove this, it is convenient to introduce two *constant* vector fields on U by

$$X = \partial_\alpha \ \ (\alpha = 1,\dots,m); \tag{4.179}$$
$$Y = a^\alpha \partial_\alpha, \tag{4.180}$$

where $(a^1,\dots,a^n)$ are the normal coordinates of some *fixed* $a \in U$. The equations

$$\tilde{\nabla}_Y Z^i = N^i W(Y); \tag{4.181}$$
$$YN^i = -\tilde{k}(Y,Z^i), \tag{4.182}$$

then hold along the geodesic $\gamma_{\vec{a}}(t)$ for $t \in [0,1]$, since there they coincide with (4.173) - (4.174).

We claim that along $\gamma_{\vec{a}}(t)$ the functions (Z^i, N^i) defined by (4.177) - (4.178) also satisfy

$$\tilde{\nabla}_Y(\tilde{\nabla}_X Z^i - N^i W(X)) = (XN^i + \tilde{k}(X,Z^i))W(Y); \tag{4.183}$$
$$Y(XN^i + \tilde{k}(X,Z^i)) = -\tilde{k}(Y, \tilde{\nabla}_X Z^i - N^i W(X)), \tag{4.184}$$

which equations are none other than (4.181) - (4.182), with the substitutions

$$Z^i \rightsquigarrow \tilde{\nabla}_X Z^i - N^i W(X); \tag{4.185}$$
$$N^i \rightsquigarrow XN^i + \tilde{k}(X,Z^i). \tag{4.186}$$

Note that the initial conditions for (4.183) - (4.184) follow from those to (4.181) - (4.182), viz.

$$\tilde{\nabla}_X Z^i(u_0) - N(u_0)^i W_{u_0}(X) = 0; \tag{4.187}$$
$$XN^i(u_0) + \tilde{k}_{u_0}(X,Z^i) = 0. \tag{4.188}$$

Indeed, by the construction of geodesic normal coordinates, *at the point* u_0, the pair (Z^i, N^i) satisfies (4.181) - (4.182) for any Y, and so in particular for X. The point now is that, (4.183) - (4.184) being a first-order system, its unique solution with initial conditions zero is zero, which by (4.185) - (4.186) shows that (Z^i, N^i) solves (4.167) - (4.168), with given initial conditions.

It remains to derive (4.183) - (4.184) from (4.181) - (4.182) and the Gauss-Codazzi equations. The argument should be familiar by now, but here we go! To derive (4.183), we compute

$$\begin{aligned}
\tilde{\nabla}_Y(\tilde{\nabla}_X Z^i - N^i W(X)) &= \tilde{\nabla}_Y \tilde{\nabla}_X Z^i - (YN^i)W(X) - N^i \tilde{\nabla}_Y(W(X)) \\
&= \tilde{\nabla}_X \tilde{\nabla}_Y Z^i - \Omega(X,Y)Z^i - (YN^i)W(X) - N^i((\tilde{\nabla}_Y W)(X) + W(\tilde{\nabla}_Y X)) \\
&= \tilde{\nabla}_X(N^i W(Y)) + \tilde{k}(X,Z^i)W(Y) - \tilde{k}(Y,Z^i)W(X) \\
&\quad - (YN^i)W(X) - N^i((\tilde{\nabla}_Y W)(X) + W(\tilde{\nabla}_Y X)) \\
&= (XN^i + \tilde{k}(X,Z^i))W(Y) + N^i(\tilde{\nabla}_X(W(Y)) - (\tilde{\nabla}_Y W)(X) - W(\tilde{\nabla}_Y X)) \\
&= (XN^i + \tilde{k}(X,Z^i))W(Y),
\end{aligned} \tag{4.189}$$

where we use (4.153) to pass to the second line and use (4.182) to cancel the term $\tilde{k}(Y,Z^i)W(X)$ on the previous line. Finally, the coefficient of N^i in the penultimate line is zero by Codazzi's equation (4.155), which emerges after using (3.65) to write $\tilde{\nabla}_X(W(Y) = (\tilde{\nabla}_X W)(Y) + W(\tilde{\nabla}_X Y)$, and noting that $W(\tilde{\nabla}_X Y) - W(\tilde{\nabla}_Y X) = W(\tilde{\nabla}_X Y - \tilde{\nabla}_Y X) = 0$ because $\tilde{\nabla}_X Y = \tilde{\nabla}_Y X$, since $\tilde{\nabla}$ is torsion-free and $[X,Y] = 0$ for the constant vector fields (4.179) - (4.180). Similarly, to derive (4.184), using eqs. (4.182), (3.65), Codazzi's (4.155), and (4.181), we compute

$$\begin{aligned}
Y(XN^i + \tilde{k}(X,Z^i)) &= XYN^i + Y\tilde{k}(X,Z^i) = -X\tilde{k}(Y,Z^i) + Y\tilde{k}(X,Z^i) \\
&= (\tilde{\nabla}_Y \tilde{k})(X,Z^i) - (\tilde{\nabla}_X \tilde{k})(Y,Z^i) + \tilde{k}(\tilde{\nabla}_Y X, Z^i) - \tilde{k}(\tilde{\nabla}_X Y, Z^i) \\
&\quad - \tilde{k}(Y, \tilde{\nabla}_X Z^i) + \tilde{k}(X, \tilde{\nabla}_Y Z^i) \\
&= -\tilde{k}(Y, \tilde{\nabla}_X Z^i) + \tilde{k}(X, N^i W(Y)) \\
&= -\tilde{k}(Y, \tilde{\nabla}_X Z^i - N^i W(X)),
\end{aligned} \tag{4.190}$$

since $\tilde{k}(X, W(Y)) = \tilde{k}(Y, W(X))$, which in coordinates is the identity $\tilde{k}_{\alpha\gamma} g^{\gamma\delta} \tilde{k}_{\delta\beta} = \tilde{k}_{\beta\gamma} g^{\gamma\delta} \tilde{k}_{\delta\alpha}$. This proves (4.184) and completes the local proof of Theorem 4.18. □

5 Geodesics and causal structure

In this chapter we introduce the *causal theory* of space-times, culminating in the key notion of *global hyperbolicity*. This theory also allows us to study local and global length-extremizing properties of geodesics, which are needed for the singularity theorems in the next chapter. The link with curvature as studied in the previous chapter is provided by the topic of the next section.

5.1 Geodesic deviation and Jacobi fields

In this section we give an interpretation of curvature through ***geodesic deviation***. This applies to both Riemannian and Lorentzian metrics and for the latter is a physical phenomenon, even a key prediction of GR. Let $U \in \mathscr{O}(\mathbb{R}^2)$ be connected and let $\gamma : U \to M$ be a family of curves: with $(s,t) \in U$ we write $\gamma_s(t) \equiv \gamma(s,t)$, regarding t as the 'time' parameter on each curve γ_s, and s as a parameter labeling the curves. Apart from the vector field tangent to $\gamma_s(t)$ along the t-flow,

$$\dot{\gamma}_s \equiv \gamma_*(\partial/\partial t) = \frac{\partial \gamma_s}{\partial t}, \tag{5.1}$$

on $\gamma(U)$, which gives the tangent vectors to each γ_s for fixed s as t "runs", we now also have a second vector field tangent to $\gamma_s(t)$ along the s-flow, i.e.,

$$\gamma_s{}' \equiv \gamma_*(\partial/\partial s) = \frac{\partial \gamma_s}{\partial s}. \tag{5.2}$$

Let ∇ be the Levi-Civita connection on TM. For any vector field Z defined on $\gamma(U)$, abbreviate

$$\nabla_s Z \equiv \nabla_{\gamma_s{}'} Z; \qquad\qquad \nabla_t Z \equiv \nabla_{\dot{\gamma}_s} Z. \tag{5.3}$$

Since $[\partial/\partial s, \partial/\partial t] = 0$ on $U \subset \mathbb{R}^2$ by standard calculus, on $\gamma(U)$ we have, cf. (4.12),

$$[\gamma_s{}', \dot{\gamma}_s] = 0. \tag{5.4}$$

Therefore, because ∇ is torsion-free we have the important identity

$$\nabla_t \gamma_s{}' = \nabla_s \dot{\gamma}_s. \tag{5.5}$$

Another application of (5.4), with (4.10), is that for any $Z \in \mathfrak{X}(\gamma(U))$ we have

$$[\nabla_t, \nabla_s] Z = \Omega(\dot{\gamma}_s, \gamma_s{}') Z. \tag{5.6}$$

Now assume that each curve $t \mapsto \gamma_s(t)$ is a geodesic, so that $\nabla_t \dot{\gamma}_s = 0$, and take $Z = \dot{\gamma}_s$. Using also (5.5), eq. (5.6) becomes the ***Jacobi equation*** or ***equation of geodesic deviation***

$$\nabla_t^2 \gamma_s{}' = \Omega(\dot{\gamma}_s, \gamma_s{}') \dot{\gamma}_s; \qquad\qquad \nabla_t^2 \left(\frac{\partial \gamma_s^\rho}{\partial s} \right) = R^\rho_{\sigma\mu\nu} \frac{\partial \gamma_s^\sigma}{\partial t} \frac{\partial \gamma_s^\mu}{\partial t} \frac{\partial \gamma_s^\nu}{\partial s}. \tag{5.7}$$

We now change perspective and start from a *single* geodesic γ. We then define a ***Jacobi field*** along γ as any vector field J, defined along γ, that satisfies Jacobi's equation

$$\nabla_t^2 J = \Omega(\dot{\gamma}, J) \dot{\gamma}; \tag{5.8}$$

$$\nabla_t^2 J^\rho = R^\rho_{\sigma\mu\nu} \frac{d\gamma^\mu}{dt} \frac{d\gamma^\sigma}{dt} J^\nu. \tag{5.9}$$

Clearly, any one-parameter family of geodesics produces a Jacobi field along any fixed one of them by the above procedure. Conversely, Jacobi fields give rise to such a family:

Proposition 5.1 *Any solution J of (5.8) or (5.9) along a geodesic γ enables one to extend γ to a one-parameter family (γ_s) of geodesics for which $\gamma = \gamma_0$ and*

$$J = \gamma_0'. \tag{5.10}$$

This will be proved in the next subsection, since we need the exponential map for the proof.

Proposition 5.2 *The collection of Jacobi fields along a given geodesic $\gamma\colon [a,b] \to M$ forms a vector space J_γ of dimension $2\dim(M)$. Specifically, one has a linear isomorphism*

$$\mathsf{J}_\gamma \cong T_{\gamma(a)}M \oplus T_{\gamma(a)}M; \tag{5.11}$$

$$J \mapsto (J(a), \nabla_t J(a)). \tag{5.12}$$

Moreover, if $J(a)$ and $\nabla_t J(a)$ are both orthogonal (parallel) to $\dot\gamma(a)$, then $J(t)$ and $\nabla_t J(t)$ remain orthogonal (parallel) to $\dot\gamma(t)$ for all $t \in [a,b]$. In the parallel case, one simply has

$$J(t) = (c_1 + (t-a)c_2)\dot\gamma(t), \tag{5.13}$$

for given initial conditions $J(a) = c_1\dot\gamma(a)$ and $\nabla_t J(a) = c_2\dot\gamma(a)$, independent of Riem.[180]

Proof. Eq. (5.8) or (5.9) is a linear second-order ODE for J, which may be rewritten as a linear first-order system $K^\rho(t) = \nabla_t J^\rho(t)$ and $\nabla_t K^\rho(t) = A^\rho_\sigma(t) J^\sigma(t)$. For such systems solutions not merely exist for short times, but for all t for which the matrix $A^\rho_\sigma(t)$ is defined. The proof of the other claims is almost trivial and is left to the reader. □

Jacobi fields play an important role in the variational properties of geodesics. To explain this we compute the second variation of the length functional (3.16) in the Riemannian case and insert the appropriate sign(s) for the Lorentzian case at the end. First, we recompute the first variation, using the powerful notion of the covariant derivative that was not yet available in §3.2. Note that, in contrast to our discussion of Jacobi fields, here we neither assume that each γ_s is a geodsic, nor (for later use in computing the second derivative) that it is parametrized by arc length (i.e. has constant speed). Using (3.65) and (3.52), (5.3), and (5.5), we obtain

$$\begin{aligned}\frac{dL(\gamma_s)}{ds} &= \int_a^b dt\, \frac{\partial}{\partial s}\sqrt{g_{\gamma_s(t)}(\dot\gamma_s(t),\dot\gamma_s(t))} \\ &= \int_a^b dt\, \frac{g_{\gamma_s(t)}(\nabla_s\dot\gamma_s(t),\dot\gamma_s(t))}{\sqrt{g_{\gamma_s(t)}(\dot\gamma_s(t),\dot\gamma_s(t))}} = \int_a^b dt\, \frac{g_{\gamma_s(t)}(\nabla_t {\gamma_s}'(t),\dot\gamma_s(t))}{\sqrt{g_{\gamma_s(t)}(\dot\gamma_s(t),\dot\gamma_s(t))}}.\end{aligned} \tag{5.14}$$

If we now do put $s=0$ (with $\gamma_0 = \gamma$) and do assume constant speed, say $\|\dot\gamma(t)\| = v$, we continue:

$$\begin{aligned}\int_a^b dt\, \frac{g_{\gamma(t)}(\nabla_t\gamma'(t),\dot\gamma(t))}{\sqrt{g_{\gamma(t)}(\dot\gamma(t),\dot\gamma(t))}} &= \frac{1}{v}\int_a^b dt\, [\partial_t(g_{\gamma(t)}(\gamma'(t),\dot\gamma(t))) - g_{\gamma(t)}(\gamma'(t),\nabla_t\dot\gamma(t))] \\ &= \frac{1}{v}\left(\Big|_b^a g(\gamma',\dot\gamma) - \int_a^b dt\; g_{\gamma(t)}(\gamma'(t),\nabla_{\dot\gamma}\dot\gamma(t))\right),\end{aligned} \tag{5.15}$$

since $\nabla_t = \nabla_{\dot\gamma}$. For fixed-endpoint variations, where $\gamma'(a) = \gamma'(b) = 0$, we therefore obtain

$$L'(\gamma) \equiv \frac{dL(\gamma_s)}{ds}(s=0) = -\frac{1}{v}\int_a^b dt\; g_{\gamma(t)}(\gamma'(t),\nabla_{\dot\gamma}\dot\gamma(t)), \tag{5.16}$$

[180] This proposition is true in the Riemannian case, and for non-null geodesics in the Lorentzian case (cf. §5.3).

since the boundary term in (5.15) vanishes. Thus we see that the extremality condition $L'(\gamma)=0$ enforces the geodesic equation (3.48), since γ' in (5.16) is arbitrary and g is nondegenerate.

We now compute the second derivative of $L(\gamma_s)$ from (5.14):

$$\begin{aligned} L''(\gamma) \equiv \frac{d^2L(\gamma_s)}{ds^2}(s=0) &= \int_a^b dt\, \frac{\partial}{\partial s}\left(\frac{g_{\gamma_s(t)}(\nabla_t {\gamma_s}'(t),\dot\gamma_s(t))}{\sqrt{g_{\gamma_s(t)}(\dot\gamma_s(t),\dot\gamma_s(t))}}\right)(s=0) \\ &= \frac{1}{v}\int_a^b dt\,[g_{\gamma_s(t)}(\nabla_s\nabla_t {\gamma_s}'(t),\dot\gamma_s(t)) + g_{\gamma_s(t)}(\nabla_t {\gamma_s}'(t),\nabla_s\dot\gamma_s(t))](s=0) \\ &\quad - \frac{1}{v^3}\int_a^b dt\,[g_{\gamma(t)}(\nabla_t\gamma'(t),\dot\gamma(t))]^2, \end{aligned} \tag{5.17}$$

where we used (5.5) to obtain the last term. Rewriting the first term using (5.6) gives

$$\begin{aligned} g_{\gamma_s}(\nabla_s\nabla_t {\gamma_s}',\dot\gamma_s)_{|s=0} &= g_\gamma([\nabla_s,\nabla_t]\gamma',\dot\gamma) + g_\gamma(\nabla_t\nabla_s\gamma',\dot\gamma) \\ &= -g_\gamma(\Omega(\dot\gamma,\gamma')\gamma',\dot\gamma) - g_\gamma(\nabla_s\gamma',\nabla_t\dot\gamma) + \frac{d}{dt}g_\gamma(\nabla_s\gamma',\dot\gamma). \end{aligned} \tag{5.18}$$

In the last line, the first term equals $-R_\gamma(\dot\gamma,\gamma',\dot\gamma,\gamma')$, the second term vanishes for geodesics, and for fixed-endpoint variations the third term also vanishes upon integration $\int_a^b dt$. Furthermore, we use (5.5), so that $g_\gamma(\nabla_t {\gamma_s}',\nabla_s\dot\gamma_s) = g_\gamma(\nabla_t {\gamma_s}',\nabla_t {\gamma_s}')$. Introducing the component

$$\gamma'_\perp = \gamma' - v^{-2}g(\gamma',\dot\gamma)\dot\gamma \tag{5.19}$$

of γ' that is perpendicular to $\dot\gamma$, we have, omitting terms containing $\nabla_t\dot\gamma \equiv \nabla_{\dot\gamma}\dot\gamma = 0$,

$$g_\gamma(\nabla_t\gamma',\nabla_t\gamma') - \frac{1}{v^2}[g_\gamma(\nabla_t\gamma',\dot\gamma)]^2 = g_\gamma(\nabla_t\gamma'_\perp,\nabla_t\gamma'_\perp). \tag{5.20}$$

Up to a boundary term vanishing upon integration for fixed-endpoint variations, we may replace the right-hand side by $-g_\gamma(\gamma'_\perp,\nabla_t^2\gamma'_\perp)$. By the symmetries of the Riemann tensor, we have

$$-\mathrm{Riem}_\gamma(\dot\gamma,\gamma',\dot\gamma,\gamma') = -\mathrm{Riem}_\gamma(\dot\gamma,\gamma'_\perp,\dot\gamma,\gamma'_\perp) = \mathrm{Riem}_\gamma(\gamma'_\perp,\dot\gamma,\dot\gamma,\gamma'_\perp) = g(\gamma'_\perp,\Omega(\dot\gamma,\gamma'_\perp)\dot\gamma),$$

so that we finally obtain ***Synge's formula*** for the second variational derivative of $L(\gamma)$:[181]

$$L''(\gamma) = -\frac{1}{v}\int_a^b dt\, g_{\gamma(t)}(\gamma'_\perp(t),\nabla_t^2\gamma'_\perp(t) - \Omega(\dot\gamma(t),\gamma'_\perp(t))\dot\gamma(t)) \tag{5.21}$$

$$= \frac{1}{v}\int_a^b dt\,[g_{\gamma(t)}(\nabla_t\gamma'_\perp(t),\nabla_t\gamma'_\perp(t)) - R(\gamma'_\perp(t),\dot\gamma(t),\gamma'_\perp(t),\dot\gamma(t))]. \tag{5.22}$$

Note that we did not assume that the curves γ_s were geodesics, except $\gamma_0\equiv\gamma$. In the Lorentzian case, for *timelike* curves,[182] one obtains *minus* the the right-hand sides in (5.21) - (5.22); this sign goes back to the one in (5.114); we invite the reader to redo the calculation for this case.

As in calculus, $L(\gamma)$ is a local minimum iff $L''(\gamma)>0$, whereas it is a local maximum iff $L''(\gamma)<0$. We will see shortly that in the Riemannian case for small t one starts out with $L''(\gamma)>0$, so that at least for short times geodesics are locally minimizing. It is clear from (5.22) that in case of negative sectional curvature this will always remain the case, but in general, $L''(\gamma)$ may go through zero. According to (5.21) and (5.8), one has $L''(\gamma)=0$ precisely when $\gamma'_\perp$ is a Jacobi field. After a necessary technical intermezzo, in §5.4 we analyze what this means.

[181] See Synge (1960), §I.6., eq. (136). It is quite remarkable that not just in the first variation (5.16), where it is expected, but also in the second variation (5.21), only the *first* s-derivative of the family γ_s appears.

[182] A curve $t\mapsto c(t)$ is *timelike* if $g_{c(t)}(\dot c(t),\dot c(t))<0$ for all t where $c(t)$ is defined. See §5.3 below.

5.2 The exponential map

Given some metric, fix $x \in M$ and define $\mathscr{V}_x \subset T_xM$ as the set of vectors $X \in T_xM$ for which the geodesic $t \mapsto \gamma_X^{(x)}(t)$ emanating at x with initial velocity X, i.e.,

$$\gamma_X^{(x)}(0) = x; \qquad \dot{\gamma}_X^{(x)}(0) = X, \tag{5.23}$$

is defined at least for $0 \leq t \leq 1$. If (M,g) is complete, then $\mathscr{V}_x = T_xM$ for all x. Note that each $\mathscr{V}_x$ is automatically open and ***star-shaped*** in that $X \in \mathscr{V}_x$ implies $tX \in \mathscr{V}_x$ for all $t \in [0,1]$. This follows because for any t for which $\gamma_X^{(x)}(t)$ is defined (for given X), and any $\rho > 0$, one has

$$\gamma_{\rho X}^{(x)}(t) = \gamma_X^{(x)}(\rho t). \tag{5.24}$$

Indeed, the left-hand side solves (3.24) with the same initial condition as the right-hand side.

The ***exponential map*** $\exp_x : \mathscr{V}_x \to M$ (based at x) is then defined by

$$\exp_x(X) = \gamma_X^{(x)}(1). \tag{5.25}$$

This map underlies many proofs and hence we take the reader through its main features.[183]

1. *At each x the set $\mathscr{V}_x \subset T_xM$ can be shrunk to an open star-shaped subset $\mathscr{U}_x \subset \mathscr{V}_x$ on which*

$$\exp_x : \mathscr{U}_x \to M$$

is a diffeomorphism onto its image U_x, called a **normal neighbourhood** *of x.* This follows from the inverse function theorem plus the observation that the derivative (pushforward) of $\exp_x$ at $0 \in T_xM$ is the identity map (as is easily verified). Moreover, *U_x may be shrunk to a* **convex neigbhourhood** *W_x of x: this means that W_x is a normal nbhd of any of its points, so that any two points of W_x may be connected by a unique geodesic.*

Eq. (5.24) then implies that $\exp_x$ maps each line segment $\{tX \mid 0 \leq t \leq 1\}$ in T_xM to the geodesic segment $\{\gamma_X^{(x)}(t) \mid 0 \leq t \leq 1\}$ in M. Conversely, *geodesics within U_x emanating from x are flattened by* $\exp_x^{-1}$. This is because any point $y = \exp_x(X) \in U_x$ is connected to x by a unique geodesic *within* U_x, viz. $\gamma_X^{(x)}$, where $X = \exp_x^{-1}(y)$ (there may be other geodesics from x to y, but if so, these leave U_x). To see this, consider some geodesic

$$c : [0,1] \to M; \qquad c(0) = x; \qquad c(1) = y, \tag{5.26}$$

and take $Y = \dot{c}(0)$. Uniqueness of geodesics c with given initial data $c(0)$ and $\dot{c}(0)$, yields

$$c(t) = \gamma_Y^{(x)}(t). \tag{5.27}$$

Then $c([0,1]) \subset U_x$ implies $Y \in \mathscr{U}_x$, and the endpoint matching condition

$$\gamma_Y^{(x)}(1) = y = \gamma_X^{(x)}(1) \tag{5.28}$$

enforces $Y = X$, which implies $c = \gamma_X^{(x)}$.

[183] O'Neill (1983), Senovilla (1998), and Minguzzi (2019) are good references for this material.

2. *Jacobi fields give the pushforward of the exponential map.* For each $X \in \mathscr{V}_x$ we have

$$(\exp_x)'_X : T_X(T_xM) \to T_{\exp_x(X)}M. \tag{5.29}$$

Identifying $T_X(T_xM) \cong T_xM$, which is done through the identification

$$Z \in T_xM \qquad \leftrightarrow \qquad \frac{d(X+tZ)}{dt}(t=0) \in T_X(T_xM), \tag{5.30}$$

(5.29) becomes a linear map $(\exp_x)'_X : T_xM \to T_{\exp_x(X)}M$. Take $Z \in T_xM$ (not necessarily orthogonal to $X = \dot{\gamma}_X^{(x)}(0)$) and let $J_Z(t)$ be the Jacobi field along $\gamma_X^{(x)}$ with boundary conditions $J(0) = 0$ and $\nabla_t J_Z(0) = Z$. For $t \in [0,1]$) we have $J_Z(t) = (\exp_{x_0})'_{tX}(tZ)$, so

$$(\exp_x)'_X(Z) = J_Z(1). \tag{5.31}$$

3. The exponential map leads to the idea of ***(geodesic) normal coordinates*** (GNC) relative to both some point $x_0 \in M$ and a choice of an orthonormal basis (e_μ) of $T_{x_0}M$. That is,

$$g_{x_0}(e_\mu, e_\nu) = \delta_{\mu\nu} \qquad \text{(Riemannian case)}; \tag{5.32}$$
$$g_{x_0}(e_\mu, e_\nu) = \eta_{\mu\nu} \qquad \text{(Lorentzian case)}. \tag{5.33}$$

These coordinates are defined on the chart U_{x_0}, as follows: the normal coordinates of $x \in U_{x_0}$ are the coordinates of $\exp_{x_0}^{-1}(x) \in T_{x_0}M$ with respect to the given basis of $T_{x_0}M$. In other words, if $X = x^\mu e_\mu$, then $x = \exp_{x_0}(X)$ has GNC x^μ, or, equivalently:

The normal coordinates (x^μ) *label the point* $\exp_{x_0}(x^\mu e_\mu) = \gamma_{x^\mu e_\mu}^{(x_0)}(1)$.

In particular, x_0 has GNC $x^\mu = 0$. By definition of $\exp_{x_0}$, and using (5.24) we also have

$$\frac{\partial f}{\partial x^\mu}(x^0) = \frac{d}{dt} f(\gamma_{te_\mu}^{(x_0)}(1))_{|t=0} = \frac{d}{dt} f(\gamma_{e_\mu}^{(x_0)}(t))_{|t=0} = e_\mu f(x_0), \tag{5.34}$$

so that in $T_{x_0}M$ we have $\partial_\mu = e_\mu$ and hence, say for the Lorentzian case, by (5.33),

$$g_{\mu\nu}(0) = g_{x_0}(e_\mu(t), e_\nu) = \eta_{\mu\nu}. \tag{5.35}$$

In GNC, by (5.24) the coordinates of the curve $t \mapsto \gamma_{x^\mu e_\mu}(t)$ are

$$x^\mu(t) = tx^\mu \tag{5.36}$$

which is clearly a geodesic. The geodesic equation (3.24) then gives

$$\Gamma^\rho_{\mu\nu}(x(t))x^\mu x^\nu = 0, \tag{5.37}$$

For $t = 0$ and $t = 1$ this gives, respectively,

$$\Gamma^\rho_{\mu\nu}(0) = 0; \tag{5.38}$$
$$\Gamma^\rho_{\mu\nu}(x)x^\mu x^\nu = 0. \tag{5.39}$$

From (4.15), the first one gives

$$\partial_\nu g_{\rho\mu} + \partial_\mu g_{\rho\nu} - \partial_\rho g_{\mu\nu} = 0. \tag{5.40}$$

Cyclic permutation of indices gives

$$\partial_\nu g_{\rho\mu} - \partial_\mu g_{\rho\nu} + \partial_\rho g_{\mu\nu} = 0. \tag{5.41}$$

Adding these equations sharpens (5.38) to

$$\partial_\rho g_{\mu\nu}(0) = 0. \tag{5.42}$$

But this is where the buck stops: for the second derivative a calculation shows that

$$\partial_\rho \partial_\sigma g_{\mu\nu}(0) = -\tfrac{1}{3}(R_{\mu\sigma\nu\rho}(0) + R_{\mu\rho\nu\sigma}(0)), \tag{5.43}$$

which is equivalent to (4.39) and, in GNC, to

$$g_{\mu\nu}(x) = \eta_{\mu\nu} - \tfrac{1}{3} R_{\mu\rho\nu\sigma} x^\rho x^\sigma + O(x^3), \tag{5.44}$$

Finally, since $f(t) = g_{\gamma^{(x_0)}(t)}(\dot\gamma^{(x_0)}(t), \dot\gamma^{(x_0)}(t))$ is constant along $\gamma_X^{(x_0)}$, i.e. in t, in GNC

$$g_{\mu\nu}(x) x^\mu x^\nu = g_{\mu\nu}(0) x^\mu x^\nu = \eta_{\mu\nu} x^\mu x^\nu, \tag{5.45}$$

as the left-hand side is $f(1)$ whilst the right-hand side is $f(0)$. This will be used shortly.

4. ***Gauss's Lemma*** (which will be used in §5.4) sharpens (5.45) to

$$g_{\mu\nu}(x) x^\mu = g_{\mu\nu}(0) x^\mu, \tag{5.46}$$

or, in coordinate-free form, for arbitrary $X \in \mathscr{V}_x$ and $Z \in T_x M$),

$$g_{\exp_x(X)}((\exp_x)'_X(X), (\exp_x)'_X(Z)) = g_x(X, Z). \tag{5.47}$$

This states that although the presence of the curvature in the right-hand side of (5.8) prevents the exponential map from being an isometry (which it is in flat space), the radial component of any vector along a geodesic preserves its length under $\exp_x$. To see that (5.47) is equivalent to (5.46), note that according to (5.36), in GNC we have

$$((\exp_x)'_X(X))^\mu = X^\mu, \tag{5.48}$$

so if we use (5.30) with $t \rightsquigarrow s$, by definition of the pushforward $(\exp_x)'_X$ we obtain

$$(\exp_x)'_X(Z) = d(\exp_x(X + sZ))_{|s=0}, \tag{5.49}$$

which in GNC gives

$$((\exp_x)'_X(Z))^\mu = Z^\mu. \tag{5.50}$$

Hence the left-hand side of (5.47) is $g_{\mu\nu}(x) X^\mu Z^\nu$, and since the right-hand side is obviously $g_{\mu\nu}(0) X^\mu Z^\nu$, we have proven the said equivalence.

To prove (5.46) and hence (5.47),[184] we note that (5.39) with (4.15) implies

$$(2g_{\mu\rho,\nu} - g_{\mu\nu,\rho})x^\mu x^\nu = 0. \tag{5.51}$$

Furthermore, taking (5.45) at arbitrary t, we have

$$g_{\mu\nu}(tx)x^\mu x^\nu = g_{\mu\nu}(0)x^\mu x^\nu, \tag{5.52}$$

whence, by taking the derivative ∂_ρ of both sides,

$$tg_{\mu\nu,\rho}(tx)x^\mu x^\nu + 2g_{\mu\rho}(tx)x^\mu = 2g_{\mu\rho}(0)x^\mu. \tag{5.53}$$

Combining (5.51) and (5.53) yields

$$\frac{d}{dt}(tg_{\mu\rho}(tx)x^\mu - tg_{\mu\rho}(0)x^\mu) = 0. \tag{5.54}$$

Hence we may evaluate the expression between brackets at $t = 1$, which gives (5.46).

Combinig (5.24), (5.31), and (5.47) gives, along the geodesic $\gamma_X^{(x)}$ (at least for $t \in [0,1]$),

$$g_{\gamma_X^{(x)}(t)}(J_X(t), J_Z(t)) = t^2 g_x(X,Z). \tag{5.55}$$

For example, on $M = \mathbb{R}^n$ with Euclidean metric (i.e. $g_{ij} = \delta_{ij}$) one simply has

$$J_Z(t) = tZ. \tag{5.56}$$

5. We now prove Proposition 5.1. Given $\gamma(t)$ and $J(t)$, let $c(s)$ be the unique geodesic with

$$c(0) = \gamma(0); \qquad c'(0) = J(0), \tag{5.57}$$

where $s \in (-\delta, \delta)$ for some $\delta > 0$, and $c'(s) = \partial c(s)/\partial s$ as usual. Then define vector fields $V(s)$ and $W(s)$ along $c(s)$ as the unique solutions of

$$\nabla_{c'} V(s) = 0; \qquad V(0) = \dot\gamma(0); \tag{5.58}$$

$$\nabla_{c'} W = 0; \qquad W(0) = \nabla_t J(0). \tag{5.59}$$

Then the following family does the job:

$$\gamma_s(t) = \exp_{c(s)}(tV(s) + stW(s)). \tag{5.60}$$

- For fixed s, this is $\gamma_s : t \mapsto \exp_{x_s}(tX_s)$, with $x_s = c(s)$ and $X_s = V(s) + sW(s)$. Now

$$\exp_{x_s}(tX_s) = \gamma_{tX_s}(1) = \gamma_{X_s}(t) \tag{5.61}$$

by (5.24), so $\gamma_s = \gamma_{X_s}$, emanating from $\gamma_s(0) = x_s$. This is surely a geodesic!

[184] Eq. (5.47) may also be proved from (5.31), cf. O'Neill (1983), Lemma 5.1 or Jost (2002), Corollary 4.2.2.

- To prove (5.10), we initially put

$$\tilde{J}(t) = \frac{\partial \gamma_s(t)}{\partial s}(s=0). \tag{5.62}$$

Then, using (5.57) - (5.61), we compute

$$\begin{aligned} \tilde{J}(0) &= \frac{\partial \exp_{c(s)}(0)}{\partial s}(s=0) = \frac{dc(s)}{ds}(s=0) = c'(0) \\ &= J(0); \end{aligned} \tag{5.63}$$

$$\begin{aligned} \nabla_t \tilde{J}(0) &= \nabla_t \frac{\partial}{\partial s}\gamma_s(t)_{|s=t=0} = \nabla_s \frac{\partial}{\partial t}\gamma_s(t)_{|s=t=0} \\ &= \nabla_{c'}(V(s)+sW(s))_{|s=0} = W(0) \\ &= \nabla_t J(0). \end{aligned} \tag{5.64}$$

Since J and $\tilde{J}$ solve the same Jacobi equation along γ, this implies $\tilde{J} = J$.

6. Finally, though not needed in what follows, the mathematical underpinning of the equivalence principle as usually conceived is given by the following extension of geodesic normal coordinates.[185] We just do the timelike Lorentzian case (which covers what is physically needed; the adaptation to the Riemannian case is obvious). Let $\gamma : (a,b) \to M$, where $a < 0 < b$, be an affinely parametrized timelike geodesic with unit speed, i.e.

$$g(\dot{\gamma},\dot{\gamma}) = -1, \tag{5.65}$$

and let (e_0, e_1, e_2, e_3) be an orthonormal frame in $T_{\gamma(0)}M$, i.e. (5.33) holds, with $e_0 = \dot{\gamma}(0)$. Parallel transport this frame along γ, i.e., the frame $(e_\mu(t))$ at $T_{\gamma(t)}M$ solves

$$\nabla_{\dot{\gamma}(t)} e_\mu(t) = 0; \qquad\qquad e_\mu(0) = e_\mu, \tag{5.66}$$

so that in particular $e_0(t) = \dot{\gamma}(t)$. The ***Fermi normal coordinates*** (x^μ) then refer to

$$(x^\mu) \leftrightarrow \exp_{\gamma(x^0)}\left(\sum_{i=1}^{3} x^i e_i)\right), \tag{5.67}$$

which defines a coordinate system in a suitable open nbhd of γ. It follows that at $\gamma(t)$ one has $\partial_\mu = e_\mu(t)$, so that similarly to the case of GNC along γ, i.e. $\forall t \in (a,b)$, one obtains

$$g_{\mu\nu}(\gamma(t)) = \eta_{\mu\nu}; \tag{5.68}$$

$$\Gamma^\rho_{\mu\nu}(\gamma(t)) = 0; \tag{5.69}$$

$$g_{\mu\nu,\rho}(\gamma(t)) = 0; \tag{5.70}$$

$$g_{00,ij}(\gamma(t)) = -2R_{0i0j}(\gamma(t)); \tag{5.71}$$

$$g_{0k,ij}(\gamma(t)) = \tfrac{2}{3}(R_{0jik}(\gamma(t)) + R_{0ijk}(\gamma(t))), \tag{5.72}$$

$$g_{lm,ij}(\gamma(t)) = -\tfrac{1}{3}(R_{iljm}(\gamma(t)) + R_{imjl}(\gamma(t))); \tag{5.73}$$

$$g_{\mu\nu,0\rho}(\gamma(t)) = 0. \tag{5.74}$$

[185] This refers to version 3(a) of the equivalence principle (which was not Einstein's), see §1.1. Fermi normal coordinates were introduced, along arbitrary curves, by Fermi (1922). See also Misner, Thorne, & Wheeler (1973), §13.6. The specialization to geodesics is taken from Manasse & Misner (1963). A similar construction even works for higher-dimensional submanifolds S (instead of geodesics), provided S carries $\dim(S)$ linearly independent vector fields each covariantly constant along S. See Schouten & Struik (1936), p. 106 and O'Raifeartaigh (1958).

5.3 Basic causal structure in Lorentzian manifolds

The following definitions are unique to Lorentzian geometry. A vector $X_x \in T_xM$ is called:[186]

- ***timelike*** if $g_x(X_x,X_x) < 0$ (so $X_x \neq 0$), and ***spacelike*** if $g_x(X_x,X_x) > 0$ or $X_x = 0$;
- ***lightlike*** if $g_x(X_x,X_x) = 0$ and $X_x \neq 0$, and ***null*** if $g_x(X_x,X_x) = 0$ (so X_x may be zero);
- ***causal*** if $g_x(X_x,X_x) \leq 0$ and $X_x \neq 0$ (i.e. X_x is either timelike or lightlike).

We denote the sets of these vectors at T_xM by $\mathscr{T}_x$, $\mathscr{S}_x$, $\mathscr{L}_x$, $\mathscr{N}_x$, and $\mathscr{C}_x$, respectively; that is,

$$\mathscr{T}_x := \{X_x \in T_xM \mid g_x(X_x,X_x) < 0\}; \tag{5.75}$$
$$\mathscr{S}_x := \{X_x \in T_xM \mid g_x(X_x,X_x) > 0 \text{ or } X_x = 0\}; \tag{5.76}$$
$$\mathscr{L}_x := \{X_x \in T_xM \mid g_x(X_x,X_x) = 0, X_x \neq 0\}; \tag{5.77}$$
$$\mathscr{N}_x := \{X_x \in T_xM \mid g_x(X_x,X_x) = 0\}; \tag{5.78}$$
$$\mathscr{C}_x := \{X_x \in T_xM \mid g_x(X_x,X_x) \leq 0, X_x \neq 0\}. \tag{5.79}$$

Diagonalizing the metric g_x to the Minkowski metric η, cf. (3.2), one sees that the set $\mathscr{T}_x$ of all timelike vectors in T_xM is disconnected, with two connected components: for any fixed $X_x \in \mathscr{T}_x$ one component $\mathscr{T}_x^+$ consist of all $Y_x \in T_xM$ such that $g_x(X_x,Y_x) < 0$, whereas the other, $\mathscr{T}_x^-$, contains all Y_x with $g_x(X_x,Y_x) > 0$. In Minkowski space-time M, for any x, taking $X_x = (1,0,0,0)$, we think of $Y_x \in \mathscr{T}_x^+$ as being future-directed (recall that $\eta = \mathrm{diag}(-1,1,1,1)$), and of $Y_x \in \mathscr{T}_x^-$ as past-directed. More generally, we call a Lorentzian manifold M ***time orientable*** if it has a global time-like vector field $T \in \mathfrak{X}(M)$, i.e., $g_x(T_x,T_x) < 0$ at each x. In that case,[187] we define

$$\mathscr{T}_x^+ := \{X_x \in \mathscr{T}_x \mid g_x(T_x,X_x) < 0\}; \qquad \mathscr{C}_x^+ := \{X_x \in \mathscr{C}_x \mid g_x(T_x,X_x) < 0\}; \tag{5.80}$$
$$\mathscr{T}_x^- := \{X_x \in \mathscr{T}_x \mid g_x(T_x,X_x) > 0\}; \qquad \mathscr{C}_x^- := \{X_x \in \mathscr{C}_x \mid g_x(T_x,X_x) > 0\}, \tag{5.81}$$

which gives a continuous choice $\mathscr{T}_x^+$ of a distinguished component of $\mathscr{T}_x$ as x varies, and similarly for causal and lightlike vectors.[188] Topologically we have, also without the $\pm$ suffix,

$$\mathscr{T}_x^{(\pm)} = \mathrm{int}(\mathscr{C}_x^{(\pm)}); \qquad \partial\mathscr{T}_x^{(\pm)} = \partial\mathscr{C}_x^{(\pm)} = \mathscr{L}_x^{(\pm)} \cup \{0\}. \tag{5.82}$$

Given a global time-like vector field T, a causal vector X_x is ***future-directed*** (fd) if $X_x \in \mathscr{C}_x^+$, and ***past-directed*** (pd) if $X_x \in \mathscr{C}_x^-$. We call $\mathscr{C}_x \cup \{0\} \subset T_xM$ the ***lightcone*** at x, with ***forward lightcone*** $\mathscr{C}_x^+$ and ***backward lightcone*** $\mathscr{V}_x^-$ (globally, there are no such things, in general Lorentzian manifolds). If M is time orientable, many choices of T will give the same $\mathscr{T}_x^+$-component of $\mathscr{T}_x$, namely any T' for which $g_x(T_x,T'_x) < 0$ for all x. If we say that $T \sim T'$ if this is the case, where both T and T' are timelike, this leaves only two equivalence classes, represented by T and $-T$. Each of these defines a ***time orientation*** of (M,g), and we see that a time-orientable Lorentzian manifold has just two possible time orientations. Since utmost generality is not our goal, we include a time orientation in our definition of a space-time:

Definition 5.3 *A* **space-time** *is a 4d connected Lorentzian manifold with time orientation.*

[186] We use the conventions of Minguzzi (2019), whom we thank for advice on this point.

[187] Counterexamples to this are quite artificial and it can be shown that every Lorentzian manifold has a double cover that is time orientable, cf. Minguzzi (2019), §1.7.

[188] Conversely, such a choice defines a global time-like vector field $T \in \mathfrak{X}(M)$ and hence a time orientation.

In both the Lorentzian and the Riemannian case, the ***"length"*** of $X_x \in T_xM$ may be defined by

$$\|X_x\| := \sqrt{|g_x(X_x, X_x)|}. \tag{5.83}$$

For spacelike vectors X_x, Y_x this "norm" satisfies the triangle equality $\|X_x + Y_x\| \leq \|X_x\| + \|Y_x\|$ as well as the Cauchy–Schwarz inequality $|g(X_x, Y_x)| \leq \|X_x\|\|Y_x\|$, with equality iff X_x and Y_x are collinear; the rule $g_x(X_x, Y_x) = \|X_x\|\|Y_x\| \cos\theta$ defines an angle θ between X_x and Y_x. However, for *causal* vectors X_x, Y_x the *opposite* Cauchy–Schwarz and triangle inequalities hold:[189]

$$|g_x(X_x, Y_x)| \geq \|X_x\| \cdot \|Y_x\|; \qquad \|X_x + Y_x\| \geq \|X_x\| + \|Y_x\|, \tag{5.84}$$

with equality iff X_x and Y_x are collinear. The *hyperbolic* angle θ between X_x and Y_x now satisfies

$$g_x(X_x, Y_x) = \mp\|X_x\|\|Y_x\| \cosh\theta, \tag{5.85}$$

with minus (plus) sign if X_x and Y_x are in the same (opposite) time cone(s).

We now define the corresponding global notions in M that replace the "infinitesimal" notions of timelike etc. in each tangent space T_xM. Properties of curves are defined through their tangent vectors: thus a curve γ is called ***(fd) timelike*** if all its tangent vectors $\dot\gamma$ are (fd) timelike, i.e. if $\dot\gamma(t) \in \mathscr{T}^{(+)}_{\gamma(t)}$ for all t, ***(fd) lightlike*** if all its tangent vectors are (fd) lightlike, ***(fd) causal*** if all its tangent vectors $\dot\gamma$ are (fd) causal, and ***spacelike*** if all its tangent vectors are spacelike. For example, in Minkowski space-time $(1,0,0,0)$ is a timelike *vector*, so that $t \mapsto \gamma(t) = (t,0,0,0)$ is a timelike *curve* (even a geodesic), since $\dot\gamma(t) = (1,0,0,0)$.[190] This terminology for curves in turn allows us to define various relations on M, of which the three most important ones are:[191]

- I^+: $(x,y) \in I^+$ or $y \in I^+(x)$ or $x \ll y$ if there is a fd *timelike* curve from x to y;
- J^+: $(x,y) \in J^+$ or $y \in J^+(x)$ or $x \leqslant y$ if there is a fd *causal* curve from x to y, or $x = y$;
- $E^+ := J^+ \backslash I^+$ (called ***horismos***): $y \in E^+(x)$ if $(x,y) \in J^+$ but $(x,y) \notin I^+$.

There are no timelike curves of zero length from x to x, so usually $(x,x) \notin I^+$, but if (M,g) admits closed timelike curves (like Gödel's space-time), then $(x,x) \in I^+$. On the other hand, $(x,x) \in J^+$ is always true by convention. As usual, we write $x < y$ if $x \leqslant y$ but $x \neq y$. There are similar relations I^-, J^-, and E^- defined by $(x,y) \in I^+$ iff $(y,x) \in I^+$, etc. This gives rise to

$$I^+(x) := \{y \in M \mid x \ll y\}; \qquad I^-(x) := \{y \in M \mid y \ll x\}; \tag{5.86}$$

$$J^+(x) := \{y \in M \mid x \leqslant y\}; \qquad J^-(x) := \{y \in M \mid y \leqslant x\}; \tag{5.87}$$

$$E^+(x) := J^+(x) \backslash I^+(x); \qquad E^-(x) := J^-(x) \backslash I^-(x). \tag{5.88}$$

In Minkowski space-time $\mathbb{R}^{3,1}$ these sets are easy to compute, with the result:

$$I^+(x) = \{y \in \mathbb{R}^4 \mid y^0 > x^0 + \|\vec y - \vec x\|\}; \tag{5.89}$$

$$J^+(x) = \{y \in \mathbb{R}^4 \mid y^0 \geq x^0 + \|\vec y - \vec x\|\}; \tag{5.90}$$

$$E^+(x) = \{y \in \mathbb{R}^4 \mid y^0 = x^0 + \|\vec y - \vec x\|\}. \tag{5.91}$$

[189] See O'Neill (1983), Proposition 5.30 and Corollary 5.31 or Minguzzi (2019), Theorems 1.2 and 1.3.

[190] In physics, timelike curves are potential trajectories of massive particles, whereas massless particles move on lightlike curves. Physical information should spread along causal curves; it will be an important task to prove this.

[191] Lemma 5.8 below implies that it does not matter if one uses smooth or piecewise smooth (or even C^1) curves.

The idea is that $J^+(x)$ is the causal future of x, consisting of all points of M that x can possibly influence (namely by signals or actions propagating with at most the speed of light).

More generally, the following subsets of M are defined causally for any subset $A \subset M$:

$$I^\pm(A) := \cup_{x\in A} I^\pm(x); \qquad J^\pm(A) := \cup_{x\in A} J^\pm(x); \tag{5.92}$$

$$E^\pm(A) := \cup_{x\in A} E^\pm(x) \qquad \mathscr{E}^\pm(A) := J^\pm(A)\backslash I^\pm(A), \tag{5.93}$$

where is should be noted that $\mathscr{E}^\pm(A) \subset E^\pm(A)$ without equality. Here are some basic facts.

Proposition 5.4 *1. The relations $I^\pm$ and $J^\pm$ are transitive (but $E^\pm$ is not).*

2. For any $A \subset M$ the set $I^\pm(A)$ is open in M; in particular, $I^\pm(x)$ is open for any $x \in M$.

3. The relations $I^\pm$ are open (i.e. $I^\pm \subset M \times M$ is open).

4. If $x \ll y$ and $y \leqslant z$, or $x \leqslant y$ and $y \ll z$, then $x \ll z$. Consequently, for any $A \subset M$,

$$I^+(A) = I^+(I^+(A)) = I^+(J^+(A)) = J^+(I^+(A)) \subset J^+(J^+(A)) = J^+(A). \tag{5.94}$$

5. For any $A \subset M$ one has the relations–with double equality in (5.96) iff $J^+(A)$ is closed–:

$$I^+(A) = \mathrm{int}(J^+(A)); \tag{5.95}$$

$$J^+(A) \subset \overline{I^+(A)} = \overline{J^+(A)}; \tag{5.96}$$

$$\partial I^+(A) = \partial J^+(A). \tag{5.97}$$

Proof. For the first claim, concatenate curves.[192] For the second, we prove that any $y \in I^+(x)$ has an open nbhd contained in $I^+(x)$. By definition there exists a timelike curve $\gamma : x \to y$. Take z on γ close enough to y that $y \in U_z$, so that $y = \exp_z(Z)$ for some timelike $Z \in T_zM$. Since the condition $g_y(Z,Z) < 0$ is open, there is an open nbhd $\mathscr{V}$ of Z in T_zM consisting of timelike vectors. Then $V = \exp_z(\mathscr{V})$ is an open nbhd of y, all whose points can be reached from z, and hence from x, by timelike curves, so that $y \in V_y \subset I^+(x)$. With $I^+(x)$, every $I^+(A)$ is open. A similar argument around x shows that if $(x,y) \in I^+$, then $V_x \times V_y \subset I$, so that I is open.

Claim 4 follows from Proposition 5.13 below: the hypothesis that there exists a causal curve from x to z via y that is initially timelike excludes case 3 of Proposition 5.13 (with $y \rightsquigarrow z$).

For (5.95), the inclusion $I^+(A) \subset \mathrm{int}(J^+(A))$ follows because $I^+(A)$ is open and is clearly contained in $J^+(A)$. Conversely, if $x \in \mathrm{int}(J^+(A))$ then by definition it has a nbhd contained in $J^+(A)$, which may be shrunk to a normal nbhd U_x. Thus there is a pd (= past-directed) timelike geodesic γ emanating from x that lies initially in U_x and contains some point $z \in U_x$; *a priori* all we know is that γ has a pd timelike tangent vector $\dot\gamma$ at x, but since $\dot\gamma$ is constant along γ this vector remains timelike, and since the condition $g(T,\dot\gamma) > 0$ for past directedness is open, it will also remain pd at least for a while. Hence $x \in I^+(z)$, so that, since $z \in J^+(A)$, we have $x \in I^+(J^+(A)) = I^+(A)$ by (5.94). This gives the inclusion $\mathrm{int}(J^+(A)) \subset I^+(A)$, and (5.95) has been proved. The proof of (5.96), which implies (5.97), is analogous but slightly more involved, and is left to the reader.[193] Eq. (5.96) implies (5.97), since for any set $\partial S = \overline{S}\backslash\mathrm{int}(S)$. □

[192] And smoothen them out, which is not necessary if piecewise smooth curves are used, cf. footnote 191.

[193] See e.g. O'Neill (1983), Lemma 14.6 (2). In this proof geodesically convex nbhds can be replaced with normal nbhds, as we have done in the previous steps of the proof, compared with e.g. Penrose (1972) and O'Neill (1983).

Against our Minkowski intuition, *the relations $J^{\pm}$ need not be closed.*[194] Nonetheless, in a normal nbhd U_x of some point x, one expects the causal relations $I^{\pm}$, $J^{\pm}$, and $E^{\pm}$ to be determined by those in T_xM. In Minkowski space-time this is even true globally: If we identify T_0M with M as usual, then (5.89) - (5.91) show that $I^+(0) = \mathscr{T}_0^+$ defined above, $J^+(0) = \mathscr{C}_0^+ \cup \{0\}$, and $E^+(0) = \mathscr{N}_0^+$, i.e. the forward lightcone from x. In general, the following theorem is a rigorous statement of the idea that "*space-time is locally Lorentz.*"

In what follows, for any nbhd U of x, the set $I_U^+(x)$ consists of all points $y \in U$ such that $x \ll y$ through a fd timelike curve *contained in* U, and similarly $I_U^-(x)$ for pd timelike curves, $J_U^{\pm}(x)$ for fd and pd causal curves, and $E_U^{\pm} = J_U^{\pm}(x) \backslash I_U^{\pm}(x)$. These sets, in which M is "reduced" to U, are not to be confused with e.g. $I^+(x) \cap U$, which is larger than or equal to $I_U^+(x)$, etc.

Theorem 5.5 *In any space-time the causal structure "near $x \in M$", i.e. in a normal nbhd $U_x = \exp_x(\mathscr{U}_x)$, is determined by its linearized version in T_xM, in the sense that:*

$$I_{U_x}^+(x) = \exp_x(\mathscr{T}_x^+ \cap \mathscr{U}_x); \tag{5.98}$$

$$J_{U_x}^+(x) = \exp_x((\mathscr{C}_x^+ \cup \{0\}) \cap \mathscr{U}_x); \tag{5.99}$$

$$E_{U_x}^+(x) = \exp_x(\mathscr{N}_x^+ \cap \mathscr{U}_x). \tag{5.100}$$

Moreover, if $c(\cdot)$ is a fd causal curve in U_x starting at x, then:

1. *If $\dot{c}(0)$ is timelike, then $c(t) \in I_{U_x}^+(x)$ for all $t > 0$ where $c(t)$ is defined.*
2. *If $c(t) \in E_{U_x}^+(x)$ for all t where $c(t)$ is defined, then c is a lightlike (pre)geodesic.*[195]
3. *Once c enters $I_{U_x}^+(x)$ (especially after a sejour on $E_{U_x}^+(x)$), it cannot leave $I_{U_x}^+(x)$.*

Finally, within U_x timelike / lightlike / causal geodesics from $x \in M$ are precisely the images under $\exp_x$ of timelike / lightlike / causal curves geodesics in T_xM starting at the zero vector.

Point 1 implies that although $y \in I_{U_x}^+(x)$ by definition means that there is a fd *timelike* curve c from x to y in U_x, i.e. $\dot{c}(t)$ is timelike *for all* t, to guarantee that $y \in I_{U_x}^+(x)$ it is enough that there is a fd *causal* curve from x to y in U_x for which *only* $\dot{c}(0)$ is timelike. Similarly, although $y \in E_{U_x}^+(x)$ by definition says that there is a fd causal curve from x to y in U_x with $y \notin I_{U_x}^+(x)$, point 2 strengthens this to $y \in E_{U_x}^+(x)$ iff there is a fd causal curve c with $c(t) \notin I_{U_x}^+(x)$ *for all* t.

The proof uses the following facts about Lorentzian metrics, which are of independent interest.

Lemma 5.6 *Let V a vector space with Lorentzian metric g and associated cones defined as in (5.75), (5.77), and (5.80), where T_xM is replaced by V and we omit the suffix x.*

1. *If $X \in \mathscr{T}^+$ and $Y \in \mathscr{C}^+$, then $g(X,Y) < 0$.*
2. *If $g(Y,Y) \leq 0$ and $g(X,Y) = 0$ for some lightlike vector X, then Y is proportional to X.*

[194] For example, remove $(t,x) = (1,1)$ from $2d$ Minkowski space-time $\mathbb{M}_2$ and look at $J^+(0,0)$: the light-ray $s \mapsto (s,s)$ from $(1,1)$ is missing. Or remove the closed horizontal line segment from $(t,x) = (2,-1)$ to $(2,1)$ from $\mathbb{M}_2$ and call this $\mathbb{M}_2'$ or ***Quinten space-time***. This removes the closed triangle with corners $(2,-1)$, $(3,0)$ and $(2,1)$ from $J^+(0,0)$ in $\mathbb{M}_2$, so that the set $J^+(0,0)$ is not closed in $\mathbb{M}_2'$.

[195] Recall that a pregeodesic is a geodesic up to reparametrization. A lightlike curve is not necessarily a pregeodesic, e.g. $t \mapsto (t, \sin t, \cos t)$ in $3d$ Minkowski space-time, and so a lightlike curve starting at x may not lie in $E_{U_x}^+(x)$.

Proof of Lemma 5.6. To prove the first claim, write $X = \lambda T + X'$ and $Y = \mu T + Y'$, where $g(T,X') = g(T,Y') = 0$ and $\lambda, \mu > 0$. Then X' and Y' are spacelike and as such satisfy the usual Cauchy–Schwarz inequality $|g(X',Y')| \leq \|X'\|\|Y'\|$. Assume $g(T,T) = -1$. Then $g(X,X) < 0$ gives $\lambda > \|X'\|$ whilst $g(Y,Y) \leq 0$ gives $\mu \geq \|Y'\|$. Thus $g(X,Y) = \lambda\mu + g(X',Y') < 0$.

The second claim follows (for example) from Lemma 4.16, since the assumption $g(Y,Y) \leq 0$ excludes the possibility that Y is spacelike, so that only the other possibility remains. □

Proof of Theorem 5.5. To ease notation we omit all reference in notation (e.g. as suffixes) to U_x and $\mathscr{U}_x$, *the restriction to which is implied throughout this proof.* We use GNC (§5.2), in which

$$c(t) = \exp_x(C(t)); \qquad c^\mu(t) = C^\mu(t), \tag{5.101}$$

where $c(t) \in M$ and $C(t) \in T_xM$. Recall that any $C \in T_xM$ gives a geodesic γ_C, which in GNC is

$$\gamma_C^\mu(t) = C^\mu t. \tag{5.102}$$

Consider a fd causal curve $c : [0,1] \to U_x$ with $c(0) = x$ and $\dot{c}(0)$ timelike, i.e.

$$g_{\mu\nu}(c(t))\dot{c}^\mu(t)\dot{c}^\nu(t) \leq 0; \tag{5.103}$$
$$g_{\mu\nu}(c(0))\dot{c}^\mu(0)\dot{c}^\nu(0) < 0. \tag{5.104}$$

Since $\dot{c}^\mu(t) = \lim_{t\downarrow 0} c^\mu(t)/t$, for sufficiently small t eq. (5.104) implies

$$g_{\mu\nu}(c(0))c^\mu(t)c^\nu(t) < 0. \tag{5.105}$$

Similarly, since $\dot{c}^\mu(t)$ is fd at $c(0)$, so is $c^\mu(t)$, for small t. Eq. (5.45) propagates (5.105) to

$$g_{\mu\nu}(c(t))c^\mu(t)c^\nu(t) < 0. \tag{5.106}$$

Furthermore, differentiating (5.105) and using Gauss's lemma in the form (5.46) gives

$$\frac{d}{dt}\left(g_{\mu\nu}(c(0))c^\mu(t)c^\nu(t)\right) = 2g_{\mu\nu}(c(0))c^\mu(t)\dot{c}^\nu(t) = 2g_{\mu\nu}(c(t))c^\mu(t)\dot{c}^\nu(t), \tag{5.107}$$

still for small t. Eqs. (5.103) and (5.106), the fact that $\dot{c}^\mu(t)$ is fd for any t by assumption, as is $\dot{c}^\mu(t)$ for small t, and Lemma 5.6.1 make the right-hand side of (5.107) negative definite, so that

$$\frac{d}{dt}\left(g_{\mu\nu}(c(0))c^\mu(t)c^\nu(t)\right) < 0, \tag{5.108}$$

for small t. Hence $g_{\mu\nu}(0)c^\mu(t)c^\nu(t)$ can only become more negative as t flows, so that (5.105), initially derived for small t, actually holds for all $t \in [0,1]$. By (5.101) and (5.33) this gives

$$\eta_{\mu\nu}C^\mu(t)C^\nu(t) < 0, \tag{5.109}$$

for all $t \in [0,1]$, or as long as the curve is in U_x. This gives $C(t) \in \mathscr{T}_x$, but since $C^\mu(t) = c^\mu(t)$ is also fd for small t and does not leave $\mathscr{T}$, by continuity we even have $C(t) \in \mathscr{T}_x^+$ for all t.

If now $y \in I^+(x)$, then by definition there is such a curve c with $c(1) = y$ (which is even timelike for all t), so that $y = \exp_x(C(1))$ with $C(1) \in \mathscr{T}_x^+$ and hence $y \in \exp(\mathscr{T}_x^+)$. This shows that $I^+(x) \subset \exp_x(\mathscr{T}_x^+)$. We now prove the converse inclusion $\exp_x(\mathscr{T}_x^+) \subset I^+(x)$.

If $y = \exp_x(C)$ for some $C \in \mathscr{T}_x^+$, then the geodesic (5.102) connects x to y by a fd timelike curve. Indeed, recall that the quantity $g_{\gamma(t)}(\dot{\gamma}(t),\dot{\gamma}(t))$ is constant in t if γ is a geodesic. Therefore,

if $\dot\gamma(0) = C$ lies in $\mathscr{T}_x^+$, then so does $\dot\gamma(t)$ for any t. Furthermore, by continuity a timelike curve cannot change its direction, since for $g_{c(t)}(\dot c(t), T_{c(t)})$ to change sign $c(\cdot)$ must leave $\mathscr{T}$.

We have now proved (5.98) as well as point 1. Point 3 follows from this, since if $y = c(t) \in I^+(x)$, one may see the remainder of $c(\cdot)$ as the continuation of a fd timelike curve starting at x, to which point 1 applies (since $I^+(x)$ is open, one may smoothen the joint just before y).

Next, one shows that each nbhd of each point $c(t)$ on a fd causal curve intersects with $\exp(\mathscr{T}_x^+)$; this is done by giving c a tiny fd timelike twist.[196] This gives $J^+(x) \subset \exp_x(\mathscr{C}_x^+)$; the converse inclusion is proved as in the timelike case. This proves (5.99).

To prove the inclusion $E^+(x) \subset \exp_x(\mathscr{N}_x^+)$, take $y \in E^+(x)$, so that there is a causal curve c from x to y, from which it follows that $C(t) \in \mathscr{C}_x^+$ for all $t \in [0,1]$. In particular, $y = \exp_x(C(1))$, where $C(1) \in \mathscr{C}_x^+$. For any $C(1) \in \mathscr{T}_x^+$ there would be a fd timelike curve (as follows from the first part of the proof), so that $C(1) \in \mathscr{N}_x^+$. Conversely, if $y = \exp_x(C)$ with $C \in \mathscr{N}_x^+$, then $y \notin I^+(x)$ by the first part of the proof, so that $y \in E^+(x)$. This proves (5.100).

To prove point 2 we show that, if $C(\cdot)$ lies in $\mathscr{C}^+$ and $c(\cdot)$ in $J^+_{U_x}(x)$, then $C(\cdot)$ lies in $\mathscr{N}_x^+$ iff $c(\cdot)$ is a lightlike geodesic (up to parametrization). From right to left this is obvious, since $c(t)$ must be $\gamma_C^{(x)}(t) = \exp_x(Ct)$ with $C \in \mathscr{N}_x^+$. For the other way round, the condition $\eta_{\mu\nu} c^\mu(t) c^\nu(t) = 0$ for C to lie in $\mathscr{N}_x^+$ implies, via (5.45) and (5.46), respectively, that

$$g_{\mu\nu}(c(t)) c^\mu(t) c^\nu(t) = 0; \qquad g_{\mu\nu}(c(t)) \dot c^\mu(t) c^\nu(t) = 0. \tag{5.110}$$

Hence by Lemma 5.6.2 the vector $\dot c^\mu(t)$ is proportional to the lightlike vector $c^\mu(t)$, so that $g_{\mu\nu}(c(t)) \dot c^\mu(t) \dot c^\nu(t) = 0$. This makes $c(\cdot)$ a lightlike *curve*; the property $\dot c^\mu \sim c^\mu$ (in GNC) makes the left-hand side of the geodesic equation (3.24) proportional to $\dot c$, so that reparametrization makes $c(\cdot)$ a lightlike *geodesic*. The final claim then restates what we knew from §5.2. □

Since $\exp_x$ is a homeomorphism and the corresponding equality holds in T_xM, we also have

$$J^+_{U_x}(x) = \overline{I^+_{U_x}(x)}. \tag{5.111}$$

We close this section with a remarkable consequence of Proposition 5.4.

Corollary 5.7 *Any* compact *space-time contains a closed fd timelike curve.*

Proof. Each set $I^+(x)$ is open by Proposition 5.4. Theorem 5.5 shows that $\cup_x I^+(x) = M$ (for any y the set $I^-(y)$ is not empty and any $x \in I^-(y)$ gives $y \in I^+(x)$). Since M is compact, $M = \cup_{i=1}^N I^+(x_i)$ for some $N < \infty$. Hence $x_1 \in I^+(x_i)$ for some i. If $i = 1$ we have $x_1 \in I^+(x_1)$ and we are ready. If not, assume $x_1 \in I^+(x_2)$, so that $x_2 \ll x_1$. Repeating this argument for the other x_i gives a chain $x_N \ll \cdots \ll x_1$. But also $x_N \in I^+(x_i)$, which gives $x_i \ll \cdots \ll x_i$. □

Compactness is sufficient but not necessary for the existence of closed fd timelike curves: for example, Gödel's space-time is topologically $\mathbb{R}^4$ and famously contains such a curve.[197] Perhaps without real justification, closed fd timelike curves are supposed not to exist in the real world. But despite Corollary 5.7, there is a lively mathematical literature on compact space-times.[198]

[196] See Senovilla (1998), Proposition 2.1 or Minguzzi (2019), Theorem 2.9 and Corollary 2.10. The difficulty of this step may be illustrated by the fact that the argument in Hawking & Ellis (1973), Proposition 4.5.1, is wrong.

[197] See Gödel (1949). For a very nice treatment of Gödel's space-time see Malament (2012), §3.1. For a very simple (spatially) noncompact example, take the ***Minkowski hypercylinder*** $M = \{(x^0, \vec x) \in \mathbb{R}^4 \mid 0 \le x^0 \le 1\} / \sim$, where $(0,\vec x) \sim (1,\vec x)$, with induced Minkowski metric. Then $I^+(x) = I^-(x) = M$ for all $x \in M$.

[198] In $d = 2$ one has interesting disanalogies between compact Lorentz surfaces and compact Riemann surfaces; for example, the uniformization theorem looks completely different in the Lorentzian case (Weinstein, 1996).

5.4 Do geodesics extremize length? Local case

Using (5.83), one may now define the length of a curve $c : [a,b] \to M$ in a Lorentzian manifold by generalizing (3.16) in the obvious way to the parametrization-independent expression

$$L(c) = \int_a^b dt\, \|\dot{c}(t)\| = \int_a^b dt\, \sqrt{|g_{c(t)}(\dot{c}(t),\dot{c}(t))|}; \qquad (c \text{ general}); \tag{5.112}$$

$$= \int_a^b dt\, \sqrt{g_{c(t)}(\dot{c}(t),\dot{c}(t))}; \qquad (c \text{ spacelike}); \tag{5.113}$$

$$= \int_a^b dt\, \sqrt{-g_{c(t)}(\dot{c}(t),\dot{c}(t))} \qquad (c \text{ timelike or causal}); \tag{5.114}$$

$$= 0 \qquad (c \text{ lightlike}), \tag{5.115}$$

The formula (3.16) for the Riemannian case is the same as the spacelike case (5.113) here. It does not matter if we work with smooth curves or with piecewise smooth curves (where in the latter case (5.112) is defined by adding the smooth pieces in the obvious way), since we have:[199]

Lemma 5.8 *If c is a piecewise smooth curve, there is a sequence (c_n) of smooth curves such that $c_n(t) \to c(t)$ and $\dot{c}_n(t) \to \dot{c}(t)$ pointwise (in the topology of M and TM, respectively), and*

$$L(c) = \lim_n L(c_n). \tag{5.116}$$

Moreover, if c is causal, then the approximating curves c_n may be chosen so as to be timelike.

Consequently, if we naively try to define a distance function on a Lorentzian manifold M by the expression that in the Riemannian case defines a proper metric in the topological sense, namely

$$d_R(x,y) := \inf\{L(c) \mid c : [a,b] \to M, c(a) = x, c(b) = y\}, \tag{5.117}$$

where the c are smooth or, equivalently, piecewise smooth curves, then $d_R(x,y) = 0$ for any $x, y \in M$. Indeed, using covers by convex nbhds one sees that any two points can be connected by a piecewise smooth lightlike curve c, which has length $L(c) = 0$, see (5.114). According to Lemma 5.8, the infimum in (5.117) remains zero if it is taken over smooth curves. In case that x and y are spacelike separated (in the sense that they can be connected by a spacelike curve), this can be remedied by stipulating that the infimum in (5.117) be taken over all spacelike curves, blocking the lightlike construction. However, if $x \leqslant y$ we have $d_R(x,y) = 0$ even if we restrict the infimum in (5.117) to piecewise smooth causal curves, or even to smooth fd timelike curves.

For causal curves, a more useful "distance" function is the so-called ***Lorentzian distance***

$$d_L(x,y) := \sup\{L(c) \mid c : [a,b] \to M, c(a) = x, c(b) = y\}, \tag{5.118}$$

defined if $x \leqslant y$, where the supremum is over all *fd causal* curves from x to y. Eq. (5.84) implies

$$d_L(x,z) \geq d_L(x,y) + d_L(y,z), \tag{5.119}$$

whenever $x \leqslant y$ and $y \leqslant z$ (which implies $x \leqslant z$), which is a reversal of the triangle inequality for a metric (Minguzzi, 2019, Theorem 2.32). Taking e.g. $x = (0,0)$, $y = (1,1)$ and $z = (0,2)$ in $2d$ Minkowski space-time shows this dramatically, since $d_L(x,z) = 2$ whilst $d_L(x,y) = d_L(y,z) = 0$, whereas with Euclidean metric d_E one has $d_E(x,z) = 2$ whilst $d_E(x,y) = d_E(y,z) = \sqrt{2}$.

[199] See Lemma 4.6.1 and Corollary 4.6.1 in Kriele (1999). Also cf. Minguzzi (2019), §2.8, and Theorem 2.37.

To sum up, we see that in the Riemannian case, where our spatial intuition comes from, a detour (notably from a geodesic) *increases* the length of a curve between two given points x and y, whereas in the Lorentzian case it *decreases* the length of a *causal* curve, assuming $y \in J^+(x)$. In particular, the critical points of the length functional on causal curves (namely causal geodesics) may be expected to *maximize* length (whereas in the Riemannian case they *minimize* length). This is indeed what happens, at least for nearby points (or small times):[200]

Proposition 5.9 *Let $x \in M$ and let $y \in U_x$ be contained in a normal nbhd of x (cf. §5.2).*

1. Riemannian *(R): x and y are connected by a unique (up to parametrization) geodesic γ of* minimal *length compared to other curves c from x to y* lying within U_x.
2. Lorentzian *(L): If $y \in I^+_{U_x}(x)$ or $y \in E^+_{U_x}(x)$, then x and y are connected by a unique (up to parametrization) fd timelike or lightlike geodesic γ, respectively, which in both cases has* maximal *length compared to other fd causal curves c from x to y in U_x.*

As in Theorem 5.5, a fd causal curve from x to $y \in E^+_{U_x}(x)$ must be a lightlike (pre)geodesic. The reason a lightlike geodesic from x to $y \in E^+_{U_x}(x)$, which has zero length, can nonetheless be of maximal length in the said way is that all other causal curves from x to y have zero length, too.

Proof. Before we discuss general Riemannian or Lorentzian manifolds, it is helpful to treat Euclidean space (R), i.e. $E = \mathbb{R}^3$ with metric $g = \mathrm{diag}(1,1,1)$ and Minkowski space (L), i.e. $\mathbb{M} = \mathbb{R}^4$ with $g = \mathrm{diag}(-1,1,1,1)$, with norm (5.83) and length (5.112). The following considerations rely on the radius function $r : M \to \mathbb{R}^+$ and the radial vector field R on M, defined by

$$r(z) := \|z\|; \tag{5.120}$$
$$R_z := z/\|z\|, \tag{5.121}$$

where we identify with T_zM with M. In $\mathbb{R}^3$ one has $R_z = \partial/\partial r$ in polar coordinates. Note that

$$g_z(R_z, R_z) = \pm 1; \tag{5.122}$$

here and in what follows the plus sign is for (R) and the minus sign applies to (L). Without loss of generality we may put $x = 0$. For any y whatsoever (R) or any y such that $x \ll y$ (L) the line $\gamma(t) = yt$ is a geodesic $\gamma : [0,1] \to M$ from x to y, with length

$$L(\gamma) = \int_0^1 dt\, \|y\| = \|y\| = r(y). \tag{5.123}$$

For (L) we first do the timelike case. Take a fd causal curve c from $x = 0$ to y. Then

$$\dot{c} = \pm g(\dot{c}, R)R + N, \tag{5.124}$$

where $g(N,R) = 0$, decomposes $\dot{c}$ into a parallel and a normal component to R. It follows that

$$\|\dot{c}\|^2 = g(\dot{c}, R)^2 \pm g(N,N), \tag{5.125}$$

with $g(N,N) \geq 0$ also in (L), where the vector R is timelike, and hence N is spacelike. Hence

$$\|\dot{c}\| \geq g(\dot{c}, R); \qquad (R) \tag{5.126}$$
$$\|\dot{c}\| \leq -g(\dot{c}, R), \qquad (L) \tag{5.127}$$

[200] This is also suggested by the twin paradox of special relativity: the twin sister leaving earth and returning cannot always travel on a geodesic and hence her curve c has shorter length (experienced by her as proper time).

since for L we have $g(\dot{c},R)<0$. The completion of the argument relies on the computation

$$\frac{d}{dt}r\circ c(t)=\frac{d}{dt}\|c(t)\|=\frac{d}{dt}\sqrt{\pm g_x(c(t),c(t))}=\pm\frac{g_x(\dot{c}(t),c(t))}{\sqrt{\pm g_x(c(t),c(t))}}=\pm g_{c(t)}(\dot{c},R). \tag{5.128}$$

Assuming $c:[0,1]\to M$ with $c(0)=x=0$ and $c(1)=y$, together with the estimates (5.126) - (5.127) and (5.123), the above computation gives

$$L(c)=\int_0^1 dt\,\|\dot{c}(t)\|\geq\int_0^1 dt\,g_{c(t)}(\dot{c}(t),R)=\Big|_0^1 r\circ c=r(y)=L(\gamma); \qquad (\mathrm{R}) \tag{5.129}$$

$$L(c)=\int_0^1 dt\,\|\dot{c}(t)\|\leq-\int_0^1 dt\,g_{c(t)}(\dot{c}(t),R)=\Big|_0^1 r\circ c=r(y)=L(\gamma), \qquad (\mathrm{L}) \tag{5.130}$$

with equalities iff $g(N,N)=0$ and hence, since N is spacelike, if $N(t)=0$ for all t. In that case, $\dot{c}$ is proportional to R and hence to c, i.e. $\dot{c}(t)=\lambda(t)c(t)/\|c(t)\|$ for some function λ. This is solved by $c(t)=zf(t)$ for some $z\in M$ and suitable f. Since $c(1)=y$ this means that $c(t)=yf(t)/f(1)$. Because $\gamma(t)=yt$, this gives $c=\gamma$ up to reparametrization, i.e. $c(t)=\gamma(s(t))$.

In general, define the radius $r:U_x\to\mathbb{R}^+$ and the radial vector field R on U_x by

$$r(\exp_x(Z)):=\|Z\|; \tag{5.131}$$

$$R_{\exp_x(Z)}:=\frac{(\exp_x)'_Z(Z)}{\|(\exp_x)'_Z(Z)\|}, \tag{5.132}$$

A curve $c:[0,1]\to M$ from x to y in U_x may be written as $c(t)=\exp_x(C(t))$. Then

$$\begin{aligned}\frac{d}{dt}r\circ c(t)&=\frac{d}{dt}\|C(t)\|=\frac{d}{dt}\sqrt{\pm g_x(C(t),C(t))}=\pm\frac{g_x(\dot{C}(t),C(t))}{\sqrt{\pm g_x(C(t),C(t))}}\\&=\pm\frac{g_{c(t)}((\exp_x)'_{C(t)}(\dot{C}(t)),(\exp_x)'_{C(t)}(C(t)))}{\sqrt{\pm g_{c(t)}((\exp_x)'_{C(t)}(C(t)),(\exp_x)'_{C(t)}(C(t)))}}=\pm g_{c(t)}(\dot{c},R),\end{aligned} \tag{5.133}$$

where we used Gauss's lemma in both the denominator and the numerator. On the other hand, "the" geodesic within U_x from x to $y=\exp_x(Y)$ is given by $\gamma_Y^{(x)}$, where

$$L\left(\gamma_Y^{(x)}\right)=\int_0^1 dt\,\|\dot{\gamma}_Y^{(x)}(t)\|=\|\dot{\gamma}_Y^{(x)}(0)\|=\|Y\|=r(y), \tag{5.134}$$

since for geodesics $\gamma=\gamma_Y^{(x)}$ the velocity $\|\dot{\gamma}(t)\|$ is t-independent. Eqs. (5.133) and (5.134) imply that the computation (5.129) - (5.130) can be repeated *verbatim*, once again yielding

$$L(c)\geq L(\gamma_Y); \qquad (R) \tag{5.135}$$

$$L(c)\leq L(\gamma_Y). \qquad (L) \tag{5.136}$$

Finally, also the proof of uniqueness of γ up to reparametrization reduces to the flat case, since the condition $\dot{c}(t)\sim R_{c(t)}$ comes down to $\dot{C}(t)\sim C(t)$. This completes the timelike case $y\in I^+_{U_x}(x)$. The case $y\in E^+_{U_x}(x)$ follows from the end of the proof of Theorem 5.5, which excludes timelike curves from x to y *and* forces the lightlike curves to be lightlike (pre)geodesics. □

5.5 Do geodesics extremize length? Global case

Being restricted to normal neighbourhoods U_x, Proposition 5.9 is local in nature; things may change beyond U_x (for given x). Here is the crucial notion, which applies to the Riemannian case in general, and to timelike or spacelike (but not: lightlike) geodesics in Lorentzian manifolds.

Definition 5.10 *A* **conjugate point** *along a geodesic $\gamma : [a,b] \to M$ is a point $\gamma(c)$, $c \in (a,b]$, for which there is a nonzero Jacobi field J along $\gamma : [a,c] \to M$ that vanishes at both a and c.*

A conjugate point is defined *relative to* $\gamma(a)$. It is independent of the parametrization of γ. The point of interest is the earliest one, if it exists. Proposition 5.2 implies that J is orthogonal to $\dot\gamma$.

Proposition 5.11 *A point z on a geodesic $\gamma : [a,b] \to M$ is conjugate iff the exponential map $\exp_{\gamma(a)}$ becomes singular at z (in that its derivative fails to be injective at the point $\gamma(c)$).*

Proof. This easily follows from (5.31); we leave the details to the reader. □

Some intuition may come from the two-sphere, where the first conjugate point along a great circle emanating from (say) the South Pole is the North Pole, at which, all of a sudden, not one unique connecting curve of minimal length exists, but infinitely many. Beyond the North Pole, the initial great circle is not even the shortest one anymore, as one may go the other way round. We know from Proposition 5.1 that J arises from a variation of γ, as in (5.10), but be aware that the boundary conditions $J(a) = 0$ and $J(c) = 0$ merely imply that the variations γ_s fix the endpoints of γ_s as $s \to 0$, so that (against the intuition from the two-sphere) the existence of J does not guarantee the existence of even one alternative geodesic from $\gamma(a)$ to $\gamma(c)$.

Nonetheless, eq. (5.22) suggests that since $L''(\gamma) = 0$ at a conjugate point, something happens to the extremization property of γ. Proposition 5.11 confirms this, as it suggests that the local analysis of Proposition 5.9 may break down. The precise situation is as follows.[201]

Theorem 5.12

1. Riemannian case: *A geodesic $\gamma : [a,b] \to M$ locally* **minimizes** *the length of curves from $\gamma(a)$ to $\gamma(b)$ iff there is* no *conjugate point on γ that lies between x and y.*

2. Lorentzian case*: A* timelike *geodesic $\gamma : [a,b] \to M$ locally* **maximizes** *the length of curves from $\gamma(a)$ to $\gamma(b)$ iff there is* no *conjugate point on γ that lies between x and y.*

The "⇐" part may be proved by remarking that, as we saw in §5.4, in the Lorentzian case timelike geodesics start out maximizing length, so that $L''(\gamma) < 0$. According to (5.21), this remains the case until a conjugate point is encountered, so if this is never the case, one will have $L''(\gamma) < 0$ forever (or at least as long as the geodesic is defined). Likewise in the R case.

For the '⇒" part, we show that the sign of $L''(\gamma)$ may indeed change once a conjugate point (at which its value is zero) has been crossed; in the L case, $L''(\gamma)$ then becomes positive, and a timelike geodesic can be constructed that is longer than the given one, whereas in the R case the opposite sign change leads to new and *shorter* geodesics between the given endpoints).[202]

Indeed, let $c \in (a,b)$, with associated Jacobi field J along $\gamma([a,c])$ for which $J(a) = 0$ and $J(c) = 0$. Then $\nabla_t J(c) \neq 0$ (since otherwise $J \equiv 0$), and by Proposition 5.1 there exists a

[201] The word 'local' here means that $\gamma([a,b])$ has a nbhd U (in M) such that γ does or does not minimize or maximize length in comparison with all curves in U, i.e. with respect to "nearby" curves only.

[202] The remainder of the proof is based on the final part of the proof of Hawking & Ellis (1973), Prop. 4.5.8. For alternative proofs see Jost (2002) Theorem 4.3.1, for the Riemannian case and O'Neill (1983), Proposition 10.10 and Theorem 10.17, Wald (1984), Theorem 9.5.3, or Minguzzi (2019), Theorem 6.16, for the Lorentzian case.

one-parameter family of geodesics (γ_s) for which $J = \gamma'_{|s=0}$. Since only the component of J that is orthogonal to $\dot\gamma$ is relevant, we can make J orthogonal to $\dot\gamma$ altogether, cf. the discussion after the statement of Proposition 5.1. Furthermore, we extend J from $\gamma([a,c])$ to $\gamma([a,b])$ by making it zero on $(c,b]$. Now find some vector field K along $\gamma\colon [a,b] \to M$ that is also orthogonal to $\dot\gamma$ and in addition satisfies the boundary conditions

$$K(a) = K(b) = 0; \tag{5.137}$$

$$g_{\gamma(a)}(\nabla_t J, K) = 0; \tag{5.138}$$

$$g_{\gamma(c)}(\nabla_t J, K) = -v. \tag{5.139}$$

This is possible, since unlike the Jacobi field J, the vector field K is not meant to satisfy any particular equation. We now take $\varepsilon > 0$ and consider the vector field

$$M = \varepsilon K + \varepsilon^{-1} J. \tag{5.140}$$

For any family of curves for which $\gamma'_{|s=0} = M$, we then compute the second variation (5.21), in which by construction $\gamma'_\perp$ is replaced by M. Since J satisfies the Jacobi equation, the term proportional to ε^{-2}, which only involves J, vanishes. The term proportional to ε^2, which only involves K, stands; call it $C\varepsilon^2$ (where C may have either sign). One of the cross terms proportional to $\varepsilon \cdot \varepsilon^{-1} = 1$, involving each of J and K linearly, vanishes by the Jacobi equation for J. In the L case to be specific (where the - sign in (5.21) has to be deleted), the other cross term contributes

$$L''(\gamma) = C\varepsilon^2 + \frac{1}{v}\int_a^c dt\, g_{\gamma(t)}(J(t), \nabla_t^2 K(t) - \Omega(\dot\gamma(t), K(t))\dot\gamma(t)). \tag{5.141}$$

Here, using (3.65) and (3.52), we have

$$g_{\gamma(t)}(J(t), \nabla_t^2 K(t)) = \frac{d}{dt}(g_{\gamma(t)}(J(t), \nabla_t K(t))) - g_{\gamma(t)}(\nabla_t J(t), \nabla_t K(t)), \tag{5.142}$$

of which the first term vanishes upon integration, as $J(a) = J(c) = 0$. The second term gives

$$-g_{\gamma(t)}(\nabla_t J(t), \nabla_t K(t)) = -\frac{d}{dt}(g_{\gamma(t)}(\nabla_t J(t), K(t))) + g_{\gamma(t)}(\nabla_t^2 J(t), K(t)), \tag{5.143}$$

whose last term combines with the curvature term in (5.141) to contribute

$$g_{\gamma(t)}(K(t), \nabla_t^2 J(t) - \Omega(\dot\gamma(t), J(t))\dot\gamma(t)),$$

which vanishes by the Jacobi equation for J (using the symmetries of the Riemann tensor R). Finally, the first term in (5.143) gives, upon integration, +1, so that overall we obtain

$$L''(\gamma) = C\varepsilon^2 + 1. \tag{5.144}$$

Whatever the sign of C, for ε small enough we can arrange $L''(\gamma) > 0$, and so, since it started out negative, the sign of $L''(\gamma)$ has changed across a conjugate point, as claimed.

It is by no means excluded that there may be other variations for which $L''(\gamma)$ remains negative; for example, by picking some K for which the sign in (5.139) is positive. All that has been proved is the existence of a family of variations for which the sign does change, which is enough to prove the theorem. A more comprehensive and systematic way to handle this situation is to introduce the *index form* for the second variation of L, which, across a conjugate point, loses its property of being negative definite (L) or positive definite (R).[203]

[203] See e.g. Jost (2002), §4.2 or O'Neill (1983), chapter 10.

Theorem 5.12 gives necessary and sufficient conditions for the existence of length-minimizing geodesics in the Riemannian case and length-maximizing timelike geodesics in the Lorentzian case, but we need to find out when these conditions are met. In the Riemannian case this is settled by the second part of the Hopf–Rinow Theorem 3.4 in §3.2. The point is that prior to this theorem we only knew that by definition a *given* geodesic extremizes the length function $c \mapsto L(c)$ defined by (3.16) compared to local variations, and that, still locally, it minimizes length until a conjugate point is encountered (and fails to to so afterwards). Theorem 3.4 is a different kind of statement: it guarantees that *some* curve between any two given points x and y exists that *globally* minimizes $L(\cdot)$, i.e., not merely compared with nearby curves from x to y, but among all curves from x to y, and then this curve must be a geodesic by definition.

There is no full Lorentzian analogue of this, and for the result that comes closest (i.e. Theorem 5.30), geodesic completeness has to be replaced by *global hyperbolicity*, a concept that is unique to Lorentzian geometry (see §5.7). But we first return to the local result Proposition 5.9.2. *A priori* there seem to be four possibilities, which one could organize into a 2×2 matrix:

- For $y \in J^+(x)$ we have (i) either $y \in I^+(x)$ or $y \notin I^+(x)$, and (ii) some fd causal curve γ from x to y either does or does not maximize $L(\cdot)$ among all fd causal curves c from x to y.

The key insight is that, as in Minkowski space-time, the second options cannot go together: although Proposition 5.9.2 itself does not hold globally, the following consequence of it does.

Proposition 5.13 *If $y \in J^+(x)$ and γ is a fd causal curve from x to y, then there are three mutually exclusive possibilities (where the reference curves c are causal and go from x to y):*

1. *$y \in I^+(x)$, and there is a timelike curve c with $L(c) > L(\gamma)$;*
2. *$y \in I^+(x)$, and γ maximizes $L(\cdot)$ among all c, so that γ is a timelike (pre)geodesic;*[204]
3. *$y \notin I^+(x)$, and γ is a lightlike (pre)geodesic that maximizes $L(\cdot)$ among all c.*

Proof (sketch).[205] Since γ lies in a compact set in M one can pick finitely many points $x_1, \ldots, x_N$ on γ (where $x_1 = x$ and $x_N = y$) and a cover of γ by pairwise overlapping normal nbhds U_{x_i}, $i = 1, \ldots, N-1$, such that $x_{i+1} \in U_{x_i}$ and U_{x_i} contains the entire segment of γ from x_i to x_{i+1}.

First, if just one segment is timelike, then $y \in I^+(x)$. To see this, assume the segment from x_k to x_{k+1} is timelike, so that $x_{k+1} \in I^+(x_k)$. If $x_{k+2} \in E^+(x_{k+1})$, then, since $I^+(x_k)$ is open, one can move x_{k+1} so as to keep the segment $x_k \to x_{k+1}$ timelike whilst making the segment $x_{k+1} \to x_{k+2}$ timelike, too. If necessary this can be repeated for all future and past points (relative to x_k), yielding a timelike curve $x \to y$. Hence for case 3 in the proposition to arise, γ must be a lightlike curve, upon which Theorem 5.5.2 shows it must be a (pre)geodesic. The only causal curves from x to $y \in E^+(x)$, then, are lightlike curves, so that γ, having length zero, trivially maximizes L over all other causal curves from x to y, since these also have zero length.[206]

[204] By definition (pre)geodesics are at least C^1, so that a curve consisting of segments of lightlike geodesics with corners is not a (lightlike) geodesic. Otherwise, any two points could be connected by a lightlike geodesic.

[205] See Minguzzi (2019), Theorems 2.20 and 2.22, for details. In its most general form the proposition is valid for *continuous* causal curves, see Definition 5.20 below, and indeed the proof is best understood in that light.

[206] This suggests that $E^+(x)$ might be a good global analogue of the fd lightcone $\mathcal{N}_x^+$ in T_xM, where $y \in E^+(x)$ *if and only if* there is a lightlike geodesic from x to y. But in general this implication is only valid from left to right, as Proposition 5.13 shows: with strong focusing of light rays (e.g. near a black hole) two points x and y may be connected by a lightlike geodesic *as well as by a timelike curve*, so that $y \in I^+(x)$ and hence $y \notin E^+(x)$.

The case distinction between 1 and 2 is then trivial, since if γ maximizes L even globally, then it certainly does so locally, so that it must be a geodesic (i.e. case 2). □

Corollary 5.14 *1. A fd causal curve from x to $y \in E^+(x)$ is a lightlike (pre)geodesic.*

2. A fd causal curve from x to $y \in J^+(x)$ that does not enter $I^+(x)$ is a lightlike (pre)geodesic.

No. 1 is case 3 of Proposition 5.13.3. This implies no. 2, whose hypothesis forces $y \in E^+(x)$. □

The following ideas will play a key role in causal theory, culminating in their relevance to the abstract theory of black holes (see chapter 10). We call a subset $S \subset M$ ***achronal*** if

$$I^+(S) \cap S = \emptyset; \qquad \Leftrightarrow \qquad I^+(S) \cap I^-(S) = \emptyset; \tag{5.145}$$

that is, if no two points of S can be connected by a timelike curve. Corollary 5.14.1 then gives:

Corollary 5.15 *Every causal curve in an achronal set is a maximizing lightlike (pre)geodesic.*[207]

At the other extreme, sets with timelike curves clearly cannot be achronal, so that the above corollary covers the situations between spacelike and timelike and therefore is not very surprising at all. The few cases where it is nontrivial include the following: if $A \subset M$, then

$$S = \partial I^+(A), \tag{5.146}$$

called an ***achronal boundary***,[208] is indeed achronal. To see this, first note the implication

$$x \in \partial I^+(A) \Rightarrow I^+(x) \subset I^+(A). \tag{5.147}$$

Indeed:

- If $y \in I^+(x)$ then $I^-(y)$ is a nbhd of x and hence $I^-(y) \cap I^+(A)$ is not empty.
- If $z \in I^-(y)$ then $y \in I^+(z)$, and if also $z \in I^+(A)$ then $y \in I^+(A)$ by transitivity of I^+.
- Hence if $x, y \in \partial I^+(A)$ satisfy $y \in I^+(x)$, then $y \in I^+(A)$, which contradicts $y \in \partial I^+(A)$, since $I^+(A)$ is open.

Corollary 5.16 *For any $A \subset M$ for which $J^+(A)$ is closed, each $x \in \partial I^+(A) \backslash A$ lies on a lightlike (pre)geodesic. Thus $\partial I^+(A) \backslash A$ is a null hypersurface, which is ruled by lightlike geodesics.*[209]

Proof. If $J^+(A)$ is closed, then $\overline{I^+(A)} = J^+(A)$ by Proposition 5.4.5. Eq. (5.97) gives

$$\partial I^+(A) = \partial J^+(A) = J^+(A) \backslash I^+(A), \tag{5.148}$$

since $I^+(A)$ is open by Proposition 5.4.2 and $J^+(A)$ is closed by assumption. Since

$$J^+(J^+(A)) = J^+(A), \tag{5.149}$$

see (5.94), there is a causal curve c through any $x \in J^+(A)$. Corollary 5.15 then applies. □

Using Lemma 5.26 below this can be shown more generally for A closed,[210] but the stated version is enough for Penrose's singularity theorem as well as various other applications.

[207] An achronal set need not contain any causal curve at all; it can be spacelike, e.g. $x^0 = c$ in $\mathbb{M}$. But a spacelike hypersurface need not be achronal either! Juts take a Lorentzian cylinder with a slowly creeping up spacelike line.

[208] More generally, an ***achronal boundary*** is a set ∂F where F is a future set. See §6.4 and §10.7.

[209] See Definition 4.15. This means that a unique lightlike geodesic passes through every point of $\partial I^+(A) \backslash A$.

[210] See Proposition 10.16 in §10.7 below, which also gives more information about the lightlike geodesics.

5.6 Properties of causal curves

Many concepts in causal theory, like the Cauchy surfaces to be studied in §5.8, as well as the singularity theorems in chapter 6, rely on some specific definitions and properties of curves.

Definition 5.17
- *A curve $c : [a,b) \to M$, where $a < b \leq \infty$, is* **future extendible** *iff $\lim_{t\uparrow b} c(t)$ exist in M. If not, c is is* **future inextendible**.[211] *Likewise for $c : (a,b] \to M$.*
- *A curve $c : (a,b] \to M$, where $-\infty \leq a < b$, is* **past extendible** *iff $\lim_{t\downarrow a} c(t)$ exists. Otherwise, c is* **past inextendible**. *Likewise for $c : (a,b) \to M$.*
- *A curve $c : (a,b) \to M$ is* **inextendible** *if it is both future and past inextendible.*

Briefly: $c : I \to M$ is inextendible if I cannot be increased *(whilst keeping c continuous).*

Although so far all curves are smooth by assumption, this definition obviously applies to curves that are merely continuous, and indeed is much more natural for that class. Since a continuous curve $c : [a,b] \to M$ is always continuously extendible to $c : [a,b+\varepsilon)$, for some $\varepsilon > 0$, only the case $[a,b)$ is interesting for future (in)extendibility. Then $c : [a,b) \to M$ is future extendible iff it has a continuous extension $c : [a,b] \to M$.[212] Likewise for past (in)extendibility at a.

Intuitively, the (future) limit $\lim_{t\uparrow b} c(t)$ may not exist for three different reasons:

1. The curve moves off to infinity. For example, for $b = \infty$ define $c : \mathbb{R} \to \mathbb{R}$ by $c(t) = t$. But this may also happen in finite time: take e.g. $c(t) = 1/t$ in $M = \mathbb{R}$, with $I = (a,0)$.

2. The would-be limit point does not exist in M. For example, take the curve

$$c : [0,1) \to \mathbb{R}\backslash\{1\}; \qquad c(t) = t. \tag{5.150}$$

3. The image of c lies in a compact set, where c continues to wander around all the different limit points of its convergent subsequences, never settling. A typical examples is the curve

$$c : [0,1) \to \mathbb{R}^2; \qquad c(t) = (t, \sin(1/(1-t)), \tag{5.151}$$

which is contained in the compact set $[0,1] \times [-1,1]$ (but has infinite arc length).

A geodesic $\gamma : (a,b) \to M$, where $a < 0 < b$, is a solution to the geodesic equation (3.24) with given initial values $\gamma(0)$ and $\dot\gamma(0)$. It is called ***future complete*** if $b = \infty$. This is not *a priori* a pointwise property, as in Definition 5.17, but nonetheless it can be shown that *solutions* γ to (3.24) whose domain is maximal are precisely geodesics that are inextendible *as curves*.[213]

Definition 5.18 *A geodesic $\gamma : (a,b) \to M$ is* **future incomplete** *iff it is future inextendible and $b < \infty$. Similarly, γ is* **past incomplete** *iff it is past inextendible and $-\infty < a$. Finally, γ is* **incomplete** *if it is either future or past incomplete, or both.*

[211] Equivalently, an ***endpoint*** of $c : [a,b) \to M$ is a point $z \in M$ such that for any nbhd U of z there is $s \in [a,b)$ such that $c(t) \in U$ for all $t \geq s$. Then c is future (in)extendible iff it has an (has no) endpoint. This criterion is especially useful if $b = \infty$. For this reason a (future/past) inextendible curve is sometimes called (future/past) ***endless***.

[212] If $b = \infty$, in order to define continuity of c we say that $U \subset [a,\infty]$ is open iff it is the complement in $[a,\infty]$ of a compact set in $[a,\infty)$. In other words, topologically $[a,\infty]$ is the one-point compactification of $[a,\infty)$.

[213] By Proposition 2.5.6 and Theorem 2.5.7 in Chruściel (2020), any fd causal/timelike geodesic has an inextendible causal/timelike extension, which is maximal as a solution to the geodesic ODE with given initial values.

Proposition 5.19 *A timelike geodesic $\gamma : (a,b) \to M$ with $-\infty < a$ is future incomplete iff it is inextendible and has finite arc length, and similarly for past incompleteness, provided $b < \infty$.*

Proof. Timelike geodesics are parametrized by arc length (up to affine reparametrizations). □

In causal theory one often needs approximations that require *continuous causal* curves. This combination of words sounds problematic, because causal properties of curves c, which so far were (piecewise) smooth by convention, were defined through their tangent vectors $\dot{c}(t)$, which require differentiability of $c(t)$. Nonetheless, the following definition makes good sense.[214]

Definition 5.20 *Let $I \subset \mathbb{R}$ be an interval, which may be (semi) closed or open, and possibly (semi) infinite. A* **continuous** *curve $c : I \to M$ is* **fd causal** *if every point $x = c(t)$ on the curve ($t \in I$) has a normal nbhd U_x (cf. §5.2) such that the unique geodesic connecting x with any later point $y \in U_x$ (with $y = c(t')$ for $t' > t$) is fd causal. Similarly for pd causal curves.*

All relevant results so far, like Proposition 5.13, are true for continuous causal curves. To analyse such curves we introduce an auxiliary *complete* Riemannian metric h on M (which always exists),[215] with associated topological metric d_h defined in the usual way, cf. (3.30). Then:

Proposition 5.21 *A continuous curve $c : I \to M$ is fd causal iff (possibly after reparametrization) it is absolutely continuous and a.e. differentiable on I with $\dot{c}$ fd causal, and, for all $[s,u] \in I$,*

$$L_h\left(c_{|[s,u]}\right) = \int_s^u dt\, \sqrt{h(\dot{c}_n(t),\dot{c}_n(t))} < \infty. \tag{5.152}$$

Proof (sketch).[216] We only sketch the inference from left to right. Take $x = c(s)$ any $y = c(u)$ close enough that they both lie in a convex set U with coordinates (x^μ) in which the metric is $ds^2 = -g_{00}dt^2 + g_{ij}dx^i dx^j$ and the interpolating geodesic has $\gamma^0(t) = t$. Since γ is causal, we have $g_{ij}\dot{\gamma}^i\dot{\gamma}^j \leq g_{00}$, and since (g_{ij}) is positive definite and U has compact closure, we have $g_{ij}x^i x^j \geq C\sum_i (x^i)^2$ for some $C > 0$. In the Euclidean distance $d(x,y)^2 = \sum_\mu |x^\mu - y^\mu|^2$ on U,

$$\begin{aligned} d(c(s),c(u))^2 = d(\gamma(s),\gamma(u))^2 &= \sum_\mu |\gamma^\mu(s) - \gamma^\mu(u)|^2 = \sum_\mu \Big|\int_s^u dt\, \dot{\gamma}^\mu(t)\Big|^2 \\ &\leq \sum_\mu \int_s^u dt\, |\dot{\gamma}^\mu(t)|^2 \leq \left(1 + \frac{g_{00}}{C}\right)^2 (u-s)^2. \end{aligned} \tag{5.153}$$

Thus c is locally Lipschitz and the claims about $\dot{c}$ follow from Rademacher's theorem. □

Since the function $u \mapsto L_h\left(c_{|[s,u]}\right)$ is strictly increasing and hence invertible, any continuous causal curve c may be parametrized by , ***h-arc length***, i.e., via one of the equivalent conditions

$$h(\dot{c}_n(t),\dot{c}_n(t)) = 1; \qquad\qquad L_h\left(c_{|[s,u]}\right) = u - s. \tag{5.154}$$

By the same token, the length functional (5.112) can be defined and is finite.

[214] We follows Minguzzi (2019). For a slightly different (locally Lipschitz) approach see Chruściel (2011, 2020).

[215] Recalling from §2.1 that our manifolds M are paracompact, a partition of unity argument gives the existence of *some* Riemannian metric $\tilde{h}$ on M (Jost, 2002, Theorem 1.4.1). If M is compact, then $\tilde{h}$ is complete, cf. Theorem 3.4. So assume M is noncompact and $\tilde{h}$ is incomplete. We follow Nomizu & Ozeki (1961). For $x \in M$, define $r(x) = \sup\{r > 0 \mid B_r(x) \text{ is compact}\}$, where $B_r(x) = \{y \in M \mid d_{\tilde{h}}(x,y) \leq r\}$. If $r = \infty$ for some x, then M is compact, so $r < \infty$. Take any smooth function $\omega : M \to \mathbb{R}$ such that $\omega(x) > 1/r(x)$. Then $h = \omega^2\tilde{h}$ is complete. In particular, any incomplete Riemannian metric can be conformally rescaled so as to become complete.

[216] For a complete proof see Theorem A.1 in Candela *et al.* (2010). See also Chruściel (2011), Theorem 2.3.2.

For later use, we now present a number of technical results, in which (M,g) is just a space-time.

Lemma 5.22 *Let a fd continuous causal curve $c:(a,b)\to M$ be parametrized proportional to h-arc length. Then $b=\infty$ iff c is future inextendible, and $a=-\infty$ iff c is past inextendible.*[217]

Approximations of curves can most naively be done pointwise, as in Lemma 5.8, but this loses even continuity. Instead, we use the auxiliary metric h to define uniform convergence.[218] Despite the square brackets, we agree that in intervals $[a,b]$ both $a=-\infty$ and $b=\infty$ are allowed.

Definition 5.23 *For curves $c_n:[a_n,b_n]\to M$ and $c:[a,b]\to M$,* **uniform convergence** *$c_n\to c$ (on compacta) means that for every compact interval $[a',b']\subset[a,b]$ there is a sequence of compact intervals $[a'_n,b'_n]\subset[a_n,b_n]$ such that the following three conditions hold, as $n\to\infty$:*

$$a'_n\to a';\qquad b'_n\to b';\qquad \sup_{t\in[a'_n,b'_n]\cap[a',b']} d_h(c_n(t),c(t))\to 0. \tag{5.155}$$

This turns out to be independent of the choice of h. Uniform convergence preserves continuity of curves, and in addition it preserves causality and (forward or backward) directedness.[219]

A very important result, to be used e.g. in the proof ofTheorem 5.30, is *upper semicontinuity of the Lorentzian length functional.* For (fd) causal curves c, $L(c)$ was defined by (5.114), i.e.

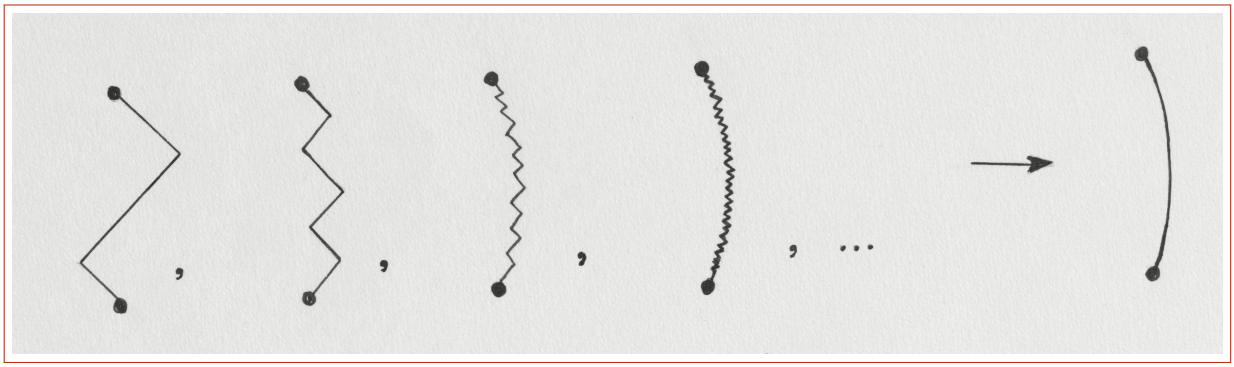
$$L(c)=\int_a^b dt\,\sqrt{-g_{c(t)}(\dot c(t),\dot c(t))}. \tag{5.156}$$

By Proposition 5.21, this expression is even defined for continuous (fd) causal curves.

Lemma 5.24 *Let c be a fd continuous causal curve $c:[a,b]\to M$, parametrized proportional to h-arc length. Then any sequence (c_n) converging uniformly to c, as in Definition 5.23, satisfies*

$$\limsup_n L(c_n)\le L(c). \tag{5.157}$$

The idea behind this lemma comes from the following figure,[220] which shows that one may *decrease* the length of a causal curve c at will by adding a chain of almost *lightlike* directions.



[217] See Minguzzi (2008), Lemma 2.6. This is even true if $c(a,b)$ is precompact, in which case c cannot have an endpoint if it is future inextendible and hence wanders around its limit points, indefinitely increasing $L_h(c)$. There is a potential ambiguity if we apply this to geodesics. These are affinely parametrized, whereas causal curves are parametrized by h-arc length. But by Proposition 2.5.6 in Chruściel (2020) a geodesic is inextendible iff it is inextendible as a causal curve (in his locally Lipschitz sense, but see footnote 216). See also footnote 213.

[218] Our discussion is based on Minguzzi (2008a, 2019). It is also possible to do such approximations without an auxiliary metric, see Hawking & Ellis (1973), Lemma 6.2.1, and O'Neill (1983), chapter 14, but this is contrived.

[219] See Lemma 2.7 in Minguzzi (2008). However, the limit curve c need not be parametrized by h-arc length.

[220] Figure redrawn from Penrose (1972), page 50, Fig. 43, by Edith de Jong.

On the other hand, *increasing* its length can only be done by adding *timelike* pieces, which becomes ever more difficult as one approaches c and hence will not lead to large length differences. Nonetheless, Lemma 5.24 is difficult to prove and we will just talk the reader through it.[221]

Lemma 5.25 *The length of any continuous causal curve c can be approximated as*

$$L(c) = \inf_{\gamma} L(\gamma), \tag{5.158}$$

where the infimum is over all interpolations of c by piecewise smooth causal geodesics γ.[222]

Intuitively, by Proposition 5.9 timelike segments of γ can only increase the length of the piece of c they interpolate, whereas lightlike segments cannot decrease it. This explains the infimum in (5.158). We may unfold the meaning of eq. (5.158) in Lemma 5.25: it states that

1. $L(c) \leq L(\gamma)$ for all piecewise smooth causal geodesics γ;
2. For any $\varepsilon > 0$ there is a γ such that $L(\gamma) \leq L(c) + \varepsilon/2$.

Applying the first point to c_n close enough to c shows that for every $\varepsilon > 0$ there is $N \in \mathbb{N}$ such that for all $n > N$ one has $L(c_n) \leq L(\gamma) + \varepsilon/2$. Hence

$$L(c_n) \leq L(\gamma) + \varepsilon/2 \leq L(c) + \varepsilon/2 + \varepsilon/2 = L(c) + \varepsilon. \tag{5.159}$$

This proves Lemma 5.24, which is one of the keys to Theorem 5.30 below on the existence of length-maximizing geodesics (which in turn leads to Hawking's singularity theorem 6.4).

Finally, we will need the following version of the ***limit curve lemma***, for which we ask the reader to first read Definition 5.27 of global hyperbolicity on the next page.[223]

Lemma 5.26 *If $(c_n : [0, b_n] \to M)$ is a sequence of fd continuous causal curves parametrized by h-arc length in a globally hyperbolic space-time such that $c_n(0) \to x$ and $c_n(b_n) \to y \neq x$, there exists a fd continuous causal curve $c : [0, b] \to M$, where $b < \infty$, as well as a subsequence of (c_n) that converges uniformly to c (cf. Definition 5.23), including $b_n \to b$ at the endpoint.*

This lemma ultimately derives from the Arzelà–Ascoli theorem. The role of global hyperbolicity is just to exclude the possibility that b_n wanders of to infinity, in which case one would have an inextendible fd causal limit curve $c : [0, \infty) \to M$ that sort of circles around y without ever reaching it,[224] cf. Lemma 5.22. Removing the assumption of global hyperbolicity allows this.

[221] Complete proofs may be found in e.g. Penrose (1972), Theorem 7.5, and Hawking & Ellis (1973), Lemma 6.7.2 (both in the setting of Theorem 5.34.1 below). We follow Minguzzi (2019), Theorems 2.37 and 2.41.

[222] This means that the smooth segments of γ are causal geodesics that have endpoints on γ and are contained in a convex nbhd that also contains the segment of γ they connect. Just think of picking sufficiently many points on c such that each adjacent pair lies in a common convex nbhd, connect this pair by a unique geodesic, and continue.

[223] Lemma 5.26 is a special case of case (i) of Theorem 2.53 in Minguzzi (2019), whose case (ii), where $b = \infty$, is excluded by our assumption of global hyperbolicity. The need for varying b_n is due to the chosen parametrization by h-arc length. In his framework of locally Lipschitz causal curves, Chruściel (2020), Proposition 2.6.2, has a simpler version of Lemma 5.26, according to which any sequence $(c : [0, \infty) \to M)$ for which $c_n(0) \to x$ and for which there is a constant $L > 0$ such that $L^{-1}|t - t'| \leq d_h(c_n(t), c_n(t')) \leq L|t - t'|$ for all n and all $t, t' \in [0, b]$ (assuming $b < \infty$), has a subsequence converging to some limit curve in the sense of Definition 5.23.

[224] Global hyperbolicity excludes compact sets $K \subset M$ that contain fd continuous causal curves $c : [0, \infty) \to K$.

5.7 Global hyperbolicity

In our context of analyzing geodesics, global hyperbolicity arises as an assumption in Theorem 5.30, which as such propagates into Theorem 6.4. It is also a key assumption in Penrose's singularity theorem. Indeed, global hyperbolicity is central to almost all of mathematical GR.[225]

Definition 5.27 *A space-time (M,g) is called:*

1. **non-imprisoning** *if no future inextendible fd causal curve c is contained in a compact subset of M (i.e., if every such curve c "eventually wanders off to infinity").*[226]

2. **globally hyperbolic** *if it is non-imprisoning and each* **double cone** *(or* **causal diamond***)*

$$J(x,y) \equiv J^+(x) \cap J^-(y) \tag{5.160}$$

is compact. N.B. by Proposition 5.4.1, $J(x,y) \neq \emptyset$ requires $x \leqslant y$, i.e. $y \in J^+(x)$.

We give various alternative characterizations of global hyperbolicity in the next section. But first, to elucidate the role of non-imprisonment we compare it to three related assumptions:[227]

Definition 5.28 *A space-time (M,g) is called:*

1. **causal** *if it contains no closed causal curves.*[228]

2. **strongly causal** *if any nbhd U_x of any $x \in M$ contains an open nbhd V_x such that any causal curve with endpoints in V_x entirely lies in V_x (as opposed to: leaving it and returning). Equivalently, if $c : I \to M$, the set $\{t \in I \mid c(t) \in V_x\}$ is connected.*

3. **non-partially imprisoning** *if there are no inextendible causal curves c that continue to return to some compact set, although they may also continue to leave it (technically: there exists no compact set $K \subset M$ for which the parameter set $c^{-1}(K) \subset \mathbb{R}$ is non-compact).*[229]

The meaning of causality should be obvious; its violation is associated with all kinds of "(murdering one's) grandfather" paradoxes. Strong causality is a form of causality stabilized against perturbations of points: there aren't even any causal curves that start at x and end at points y arbitrarily closely *near* x, except the very short direct causal curves from x to y (if $y \in J^{\pm}(x)$).

Partial imprisonment is a weakening of imprisonment, and the logical implications are:[230]

$$\textit{strongly causal} \Rightarrow \textit{non-partially-imprisoning} \Rightarrow \textit{non-imprisoning} \Rightarrow \textit{causal}.$$

[225] Our definition of global hyperbolicity (which simplifies some arguments in §5.8) is equivalent to the usual one in which non-imprisonment is replaced by strong causality. See Minguzzi (2019), Proof after Definition 4.117.

[226] This is equivalent to the same condition with future/fd changed into past/pd, see Minguzzi (2019), page 119.

[227] See Minguzzi (2008b, 2019). This is part of the ***causal ladder*** (Minguzzi & Sánchez, 2008). Imprisonmment (under the name of ***total imprisonment***) and partial imprisonment were introduced by Carter (1971a).

[228] One also says that (M,g) is ***chronological*** if there are no closed *timelike* curves.

[229] For an imprisoned curve c, the set $c^{-1}(K)$ is by definition the entire parameter space I.

[230] The first implication is Proposition 4.80 in Minguzzi (2019), the second is trivial from the definitions, and the third is Proposition 4.37 in Minguzzi (2019). The implication *strongly causal* $\Rightarrow$ *non imprisoning* is also obvious in the contrapositive; for any inextendible fd causal curve $c : [a,b) \to K$ contained in a compact set K has a limit point x as $t \to b$, but it has no endpoint, and so any limit point provides a counterexample to the definition of strong causality. See Hawking & Ellis (1973), p. 195 or Minguzzi (2019, §4.3.1) for a (contrived) example of a (totally) imprisoning space-time; but all examples must be pretty pathological.

Compactness of $J(x,y)$ holds in Minkowski space-time. In curved space-times it allows "interesting" singularities but blocks "trivial" ones, as in the case where $J(x,y)$ fails to be compact by missing a point in its interior. In that case, there is a fd causal curve from x that disappears into this point. This curve lies in the past of y and hence is visible for an observer at y, making the hole a (locally) *naked singularity*. Global hyperbolicity prevents this possibility. See §10.4.

Technically, the following equivalences will be useful (proved by elementary topology):[231]

Lemma 5.29 *Let (M,g) be a space-time. The following properties are equivalent:*

1. *Each double cone $J^+(x)\cap J^-(y)$ is closed.*
2. *All sets $J^\pm(x)$ are closed.*
3. *All sets $J^\pm(K)$, where $K\subset M$ is compact, are closed.*

Also, (M,g) is globally hyperbolic iff $J^+(K)\cap J^-(L)$ is compact for any compact $K,L\subset M$.

The following key result is independently due to Avez and Hawking.[232] To see the need for global hyperbolicity in this, note that in Minkowski space-time with the origin removed, no point $(x^0,\vec{0})$ with $x^0<0$ can be connected to $(x^0,\vec{0})$ with $x^0>0$ by a geodesic at all.

Theorem 5.30 *If (M,g) is globally hyperbolic, then any $x\in M$ and $y\in J^+(x)$ can be connected by a fd causal geodesic of finite length, which length is maximal among all fd causal curves from x to y. If $y\in I^+(x)$ then this maximizing geodesic is timelike, and if $y\in E^+(x)$ it is lightlike.*

Proof. All curves are continuous fd causal. Recall (5.117) and (5.118). Since $L(\gamma)$ and hence $d_L(x,y)$ are parameter-independent, we may assume all our curves to be parametrized by h-arc length. We also assume that all curves start at $t=0$ with $c(0)=x$. All curves c from x to y lie in $J(x,y)$, which is compact by assumption. In $J(x,y)$ we may choose our auxiliary Riemannian metric h such that $\|X\|_g\le\|X\|_h$ for all causal vectors X. Lemma 5.31 below then gives:[233]

$$d_L(x,y)\le d_R(x,y)<\infty. \tag{5.161}$$

Now take a continuous fd causal sequence (c_n) for which $\sup_n L(c_n)=d_L(x,y)$, and hence also $\limsup_n L(c_n)=d_L(x,y)$. By Lemma 5.26 this sequence has a limit curve $c:[0,b]\to M$ with $c(b)=y$. Lemma 5.24 gives

$$L(c)\le d_L(x,y)=\sup_n L(c_n)=\limsup_n L(c_n)\le L(c), \tag{5.162}$$

so that $L(c)=d_L(x,y)$. Hence c achieves the supremum in (5.118) and has maximal length. Proposition 5.13 then makes c a (smooth) causal pregeodesic with the claimed properties,[234] predicated on $y\in I^+(x)$ or $y\in E^+(x)$. Finally, reparametrization turns it into a geodesic. □

Lemma 5.31 *If (M,g) is non-imprisoning, then for any compact subset $K\subset M$ there is a constant $0<C_K<\infty$ such that for any continuous (fd) causal curve $c:[0,b]\to K$ or $c:[0,b)\to K$, parametrized by h-arc length, one has the uniform bound $L_h(c)<C_K$. In particular, $b<\infty$.*

[231] See Minguzzi (2019), Theorem 4.12, for part 1, and Galloway (2014), Proposition 4.3, for the last claim.
[232] See Avex (1963) and Hawking (1966/2014), his Adams Prize Essay, expanded into Hawking & Ellis (1973).
[233] See Minguzzi (2019), Theorem 2.55.
[234] This also follows from Theorem 2.20 in Minguzzi (2019), which states that each maximizing causal curve is a causal geodesic. As in Proposition 5.13, the proof is done by localization to convex nbhds.

5.8 Cauchy surfaces and Cauchy horizons

In this section we give various equivalent characterizations of global hyperbolicity.[235] These look quite different from each other, which is very useful in both causal and PDE theory and also illustrates the richness of the concept. First, for $x \leqslant y$ (i.e. $x \in M$ and $y \in J^+(x)$), define $C(x,y)$ as the space of continuous fd causal curves c from x to y that are defined on $I = [0,1]$ and are parametrized *proportional to* h-arc length, where h is a complete Riemannian metric, as in the previous section: hence $h(\dot{c}_n(t), \dot{c}_n(t))$ is constant a.e. in t, though not necessarily 1. Then

$$d(c_1,c_2) := \sup_{t\in[0,1]} d_h(c_1(t),c_2(t)) + |L_h(c_1) - L_h(c_2)| \tag{5.163}$$

turns $C(x,y)$ into a metric space; the first term makes the *evaluation map*

$$\mathrm{ev} : C(x,y) \times [0,1] \to M; \qquad \mathrm{ev}(c,t) = c(t) \tag{5.164}$$

continuous, upon which the second term also makes the Riemannian curve *length functional*

$$L_h(c) = \int_0^1 dt \sqrt{h(\dot{c}(t),\dot{c}(t))} \tag{5.165}$$

continuous.[236] Leray's original definition of global hyperbolicity was essentially that each space $C(x,y)$ be precompact in the compact-open topology borrowed from $C([0,1],M)$.[237] Surprisingly, this is equivalent to the existence of a *Cauchy surface*, which we define now.

[235] See Leray (1953), and Choquet-Bruhat (2014) for some history. See also Choquet-Bruhat (2009), chapter XII. Our approach combines elements of Choquet-Bruhat, *loc. cit.*, with Theorem 4.1 in Sánchez (2007).

[236] This makes convergence in the metric d stricter than uniform convergence in d_h (as in Definition 5.23, in which L_h is generally merely lower semicontinuous, analogously to upper semicontinuity of the Lorentzian length L). The metric (5.163) was introduced by Bott & Mather (1968, p. 474) and is also used by Choquet-Bruhat (2009), §XII.8.2 (who works with the class of *rectifiable* continuous causal curves). Since $I = [0,1]$ is compact, the first term in (5.163) gives the compact-open topology, see Clarke (1993), §6.2.2. One may wonder why things are so complicated. The reason is that if one allows arbitrary parametrizations of curves all hope of compactness of $C(x,y)$ or its closure are gone. The approach chosen in the main text (following Bott, Choquet-Bruhat and Sánchez) is one way around this problem by introducing preferred parametrizations, at the cost though of the unusual metric (5.163). Alternatively, Penrose (1972) and Hawking & Ellis (1973) work with the space $\hat{C}(x,y)$ of continuous fd causal curves *up to reparametrization*, i.e. one uses the *image* $c([0,1])$ in M rather than the *function* $c : [0,1] \to M$. This image space is topologized by letting any open nbhd of c (more precisely, its image in M) consist of all fd causal curves γ whose image lies in some open nbhd of $c([0,1])$ in M. This topology is very natural and coincides with the quotient of the compact-open topology on $C([0,1],M)$ to the image space, see again Clarke (1993), §6.2.2. However, unlike the approach in the main text, this procedure hardly makes sense when (M,g) is not causal, since in that case loops traversed any number of times are identified (since they have the same image in M), although they are clearly different things. Thus one assumes causality from the outset, indeed even strong causality, in which case $\hat{C}(x,y)$ need not be completed and global hyperbolicity is characterized by compactness of $\hat{C}(x,y)$. See Penrose (1972), §6 or Hawking & Ellis (1973), Proposition 6.6.2. Furthermore, Lemma 5.24 remains valid, *mutatis mutandis*, in that L is upper semicontinuous, i.e. for each $c \in C(x,y)$ and each $\varepsilon > 0$ there is a nbhd Γ of c such that $L(\gamma) \leq L(c) + \varepsilon$ for all $\gamma \in \Gamma$. See Penrose (1972), Theorem 7.5 or Hawking & Ellis (1973), Lemma 6.7.2.

[237] The topology induced by the metric (5.163) is not defined on all of $C([0,1],M)$ because of the second term; the first term would recover the compact-open topology. Restricted to $C(x,y)$, the metric topology given by (5.163) is of course finer than the compact-open topology, so that $\overline{C}(x,y)$ is also complete in the metric (5.163), and indeed $\overline{C}(x,y)$ is a good model of the abstract (Cauchy) completion of $C(x,y)$ defined for any metric space. This is necessary because despite Lemma 5.26, the space $C(x,y)$ is not itself complete in the metric (5.163), since, as already mentioned in footnote 217), uniform limits of sequences of continuous causal curves parametrized proportional to h-arc length need to be parametrized in that way and hence may disappear from $C(x,y)$.

Definition 5.32 *A* **Cauchy (hyper)surface** *in a space-time (M,g) is a subset $\Sigma \subset M$ with the property that each inextendible timelike curve intersects Σ in exactly one point.*

An easy example is the $x^0 = 0$ hypersurface in Minkowski space, which may be tilted or curved, as long as all tangent vectors remain spacelike. But neither the hyperboloid H^3_ρ defined in (4.88) nor even the forward lightcone $\partial I^+(x)$ is a Cauchy surface. Here are some first results:

Theorem 5.33 *Let (M,g) be a space-time with Cauchy surface $\Sigma \subset M$. Then:*

1. *Σ is a closed connected achronal 3d topological submanifold of M.*
2. *Any other possible Cauchy surface in M is homeomorphic to Σ.*
3. *M is homeomorphic to $\mathbb{R} \times \Sigma$.*

We will give more precise results Theorem 5.44, especially concerning the possible smoothness of all constructions and the existence of spacelike Cauchy surfaces, but for a first acquaintance with Cauchy surfaces the above facts are enough (and indeed historically they were the first to be established).[238] The first claim is technical,[239] except for achronality which is trivial, but the second and third are fairly intuitive. If Σ and Σ' are both Cauchy surfaces, then any inextendible timelike curve meets each of them once. In particular, the integral curves of a complete timelike vector field T on M, such as the one defining its time-orientation,[240] give an identification of Σ and Σ'. Similarly, since the integral curves c of T are topologically $\mathbb{R}$ (see Lemma 5.22, extended also in the backward direction), we obtain a map

$$\mathbb{R} \times \Sigma \to M; \qquad (t, \sigma = c(0)) \mapsto c(t), \tag{5.166}$$

that is, Σ is moved (forward or backward in time) with the flow of T. This map is a bijection by definition of a Cauchy surface, and can be shown to be a homeomorphism, like the above bijection $\Sigma \cong \Sigma'$. Such arguments can be made rigorous once we have time functions (see §5.9).

Theorem 5.34 *Each of the following conditions is equivalent to global hyperbolicity of (M,g):*

1. *The space $C(x,y)$ is precompact for all $x \leqslant y$ (i.e. its closure $\overline{C}(x,y)$ is compact).*
2. *For each $x \leqslant y$ there is a constant $K_{x,y} < \infty$ such that for all $c \in C(x,y)$ one has*
$$L_h(c) < K_{x,y}. \tag{5.167}$$
3. *M has a Cauchy surface.*

A detailed proof of this theorem takes many pages and is hardly instructive, except for explaining how the various assumptions are related to each other. Short of giving a complete proof, our goal is therefore merely to sketch these relations, and refer those who want more to the literature.[241]

[238] The theory of Cauchy surfaces in GR was initiated by Geroch (1970) in the topological setting; see also Hawking & Ellis (1973), chapter 6 and O'Neill (1983), chapter 14. This theory was extended to the smooth case by Bernal and Sánchez (2003, 2005, 2006a); see also the reviews Sánchez (2005, 2007). See the end of §5.9.

[239] See e.g. O'Neill (1983), Lemma 14.29 to Corollary 14.32. Claim 3 requires a version of the limit curve lemma.

[240] If the given T is not complete, then $T/\|T\|_h$ is complete, where h is a complete Riemannian metric as usual.

[241] See for example Geroch (1970), Penrose (1972), chapters 6 and 7, Hawking & Ellis (1973), chapter 6, O'Neill

For the implication $1 \to 2$ we note that L_h is continuous and hence Weierstrass's theorem guarantees that L_h assumes a maximum value M on the compact set $\overline{C}(x,y)$. Any $K_{x,y} > M$ then satisfies (5.167). Conversely, by the Arzelà–Ascoli theorem $\overline{C}(x,y)$ is compact iff:

1. Each set $\{c(t) \mid c \in \overline{C}(x,y)\} \subset M$, where $t \in (0,1)$, is bounded;
2. The family of curves $\overline{C}(x,y)$ is equicontinuous, i.e., for each $t \in [0,1]$ and each $\varepsilon > 0$ there is $\delta > 0$ such that if $|s-t| < \delta$, then $d(c(s),c(t)) < \varepsilon$ for all $c \in \overline{C}(x,y)$.

Eq. (5.167) implies both conditions: indeed, if this is the case, then the inequalities

$$d_h(x,c(t)) \leq L_h(c) < K_{x,y} \tag{5.168}$$

make the set $\{c(t) \mid c \in \overline{C}(x,y)\}$ in clause 1 of the Arzelà–Ascoli theorem bounded. Assuming for now c is parametrized by h-arc length, we have $L_h(c(s,t)) = L_h(c)|s-t|$, and hence

$$d_h(c(s),c(t)) \leq L_h(c)|s-t| < K_{x,y}, \tag{5.169}$$

which proves equicontinuity. Hence $\overline{C}(x,y)$ is compact and we are ready with $1 \leftrightarrow 2$.

To prove that 1 or 2 is equivalent to global hyperbolicity as in Definition 5.27, first note that if $\overline{C}(x,y)$ is compact, then so is $J(x,y)$. This follows from the continuity of the evaluation map. The inequality (5.167) forces non-imprisonment by contradiction: if $K \subset M$ is compact and contains an inextendible fd continuous causal curve c, then this curve is also contained in a double cone $J(x,y)$, just proven compact, to which Lemma 5.22 and (5.167) apply.[242]

Conversely, the implication from Definition 5.27 to (5.167), immediately follows by taking $K = J(x,y)$ in Lemma 5.31. Thus Definition 5.27 and properties 1 and 2 in Theorem 5.34 are closely related and easily transferable into each other; the only technical tool was Arzelà–Ascoli.

Property 3 is quite different, and the proof of equivalence uses a whole new arsenal of techniques, each of which is also of independent interest and has many other applications in GR.

First, the analysis of Cauchy surfaces in property 3 involves the following concept:[243]

Definition 5.35 *Let $S \subset M$ be an achronal susbet of M.*

1. *The* **domain of dependence** *or* **future Cauchy development** $D^+(S)$ *of S is the set of all $x \in M$ for which every past-inextendible pd causal curve starting from x intersects S.*
2. *The* **domain of influencedomain of influence** *or* **past Cauchy development** $D^-(S)$ *of S is the set of all $x \in M$ for which every future-inextendible fd causal curve starting from x intersects S.*
3. *The* **total domain of dependence** *or* **two-sided Cauchy development** *of S is*

$$D(S) := D^+(S) \cup D^-(S). \tag{5.170}$$

(1983), chapter 14, Beem, Ehrlich, & Easley (1996), chapter 3, Kriele (1999), chapter 8, Choquet-Bruhat (2009), chapter XII, Chruściel (2011), and Minguzzi (2019), chapter 3.

[242] If $c : [0,\infty) \to K$ is the curve in question, then take $x = c(0)$ and $y \in \cap_{t \geq 0} \overline{c(t,\infty)}$, which is nonempty and lies in K by Minguzzi (2019), Proposition 2.72. The curve then lies in $\overline{C}(x,y)$.

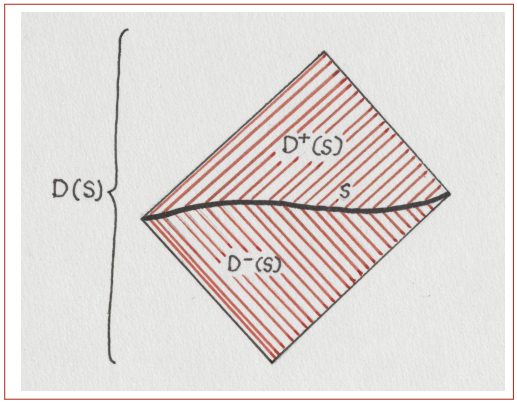
[243] Definition 5.35 makes sense for any $S \subset M$ but is only used when S is achronal, i.e. if no two points of S can be connected by a timelike curve; see (5.145). Such surfaces S carry initial data for hyperbolic PDEs and the idea is that in relativistic physics everything happening at $x \in D^+(S)$ is determined by the state of affairs at S. This is not really a theorem of mathematical physics, but it is a principle that is backed by the theory of hyperbolic PDEs. See Courant & Hilbert (1962), Choquet-Bruhat (2009), Bär, Ginoux, and Pfäffle (2007), and Earman (1995, 2007).

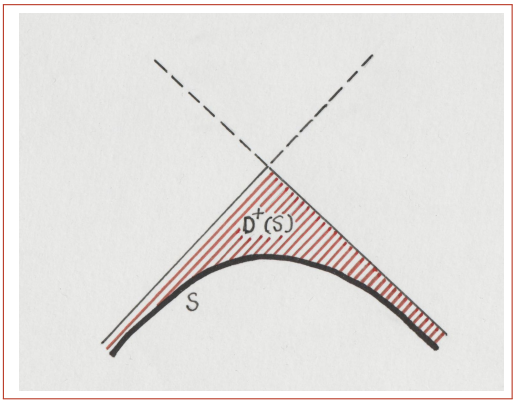
By definition, one then has the (not necessarily strict) inclusions

$$S \subset D^{\pm}(S) \subset J^{\pm}(S). \tag{5.171}$$

A simple case is $S = \{x \mid x^0 = 0\}$ in Minkowski space-time, which gives $D^+(S) = \{x \mid x^0 \geq 0\}$.

Here are four more instructive examples, all due to Penrose himself.[244]





Examples of domains of dependence and influence, taking place in 2d Minkowski space-time $(\mathbb{M}_2, \eta)$.

- *In the figure on the left, S is a generic closed, achronal, and* bounded *(and hence compact) (hyper)surface. The domains* $D^{\pm}(S)$ *are compact, too, and so is, of course, their union* $D(S)$.
- *In the figure on the right, S is the closed and achronal, but* unbounded *"southern" hyperboloid*

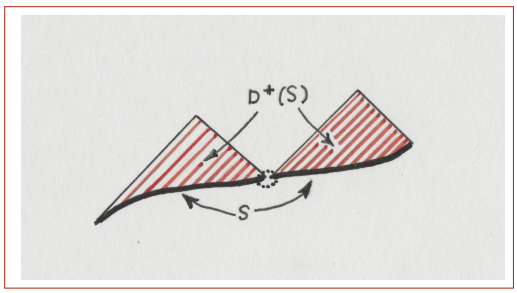
$$S = H^1_1 := \{(t,x) \in \mathbb{R}^2 \mid t = -\sqrt{x^2+1}\}, \tag{5.172}$$

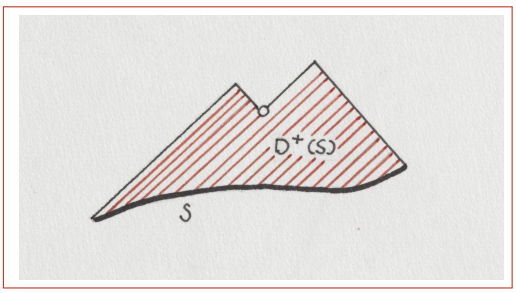
cf. (4.88)*, which asymptotes towards the past lightcone* $t = -|x|$*. For the domains* $D^{\pm}(S)$ *we find*

$$D^+(S) = \{(t,x) \in \mathbb{R}^2 \mid -\sqrt{x^2+1} \leq t < -|x|\}; \tag{5.173}$$

$$D^-(S) = \{(t,x) \in \mathbb{R}^2 \mid t \leq -\sqrt{x^2+1}\}. \tag{5.174}$$

Note that $D^+(S)$ *is not closed, since lightlike (hence causal) curves on* $t = -|x|$ *do not meet S.*





- *In the figure on the left, a point has been removed from S, which has a drastic effect on* $D^+(S)$.
- *In the figure on the right, a point has been removed from* $\mathbb{M}_2$*, with a similar effect on* $D^+(S)$.

[244] Adapted from Penrose (1972), pp. 39–40, Fig. 31–34, redrawn by Edith de Jong. Note that Penrose defines the domains $D^{\pm}(S)$ using timelike curves instead of causal curves. If we write these as $D^{\pm}_P(S)$, then for closed achronal sets S one has $D^{\pm}_P(S) = \overline{D^{\pm}(S)}$ (Minguzzi, 2019, Proposition 3.10), so that $D^{\pm}_P(S)$, unlike $D^{\pm}(S)$, is closed.

In any dimension $d \geq 2$, the hyperboloid example (5.172), adding the "north", becomes:

$$S = \pm H_1^3 = \left\{ x \in \mathbb{R}^4 \mid x^0 = \pm\sqrt{|\vec{x}|^2+1} \right\}; \tag{5.175}$$

$$D^\pm(\pm H_1^3) = J^\pm(\pm H_1^3) = \left\{ x \in \mathbb{R}^4 \mid \pm x^0 \geq \sqrt{|\vec{x}|^2+1} \right\}; \tag{5.176}$$

$$D^\pm(\mp H_1^3) = J^\pm(\mp H_1^3) \cap I^\mp(0) = \left\{ x \in \mathbb{R}^4 \mid -|\vec{x}| < \pm x^0 \leq \sqrt{|\vec{x}|^2+1} \right\}. \tag{5.177}$$

Following Hawking, we now define the ***future/past Cauchy horizons*** $H_C^{+/-}(S)$ of S by

$$H_C^+(S) := \overline{D^+(S)} \backslash I^-(D^+(S)) = \{x \in \overline{D^+(S)} \mid I^+(x) \cap \overline{D^+(S)} = \emptyset\}; \tag{5.178}$$

$$H_C^-(S) := \overline{D^-(S)} \backslash I^+(D^-(S)) = \{x \in \overline{D^-(S)} \mid I^-(x) \cap \overline{D^-(S)} = \emptyset\}; \tag{5.179}$$

$$H_C(S) := H_C^+(S) \cup H_C^-(S). \tag{5.180}$$

That is, $H_C^+(S)$ consists of all points $x \in D^+(S)$ that precede no other point in $D^+(S)$, etc.[245] As any point beyond $H_C^+(S)$ can be influenced by events outside S (etc.), the Cauchy horizons $H_C^\pm(S)$ measure the failure of S to be a Cauchy surface, cf. Proposition 5.38 below. But first, we simplify eqs. (5.178) - (5.179) under further assumptions on S (beyond it being achronal).

Definition 5.36 *1. $S \subset M$ is* **acausal** *if there is no causal curve that starts and ends at S.*

2. The **edge** *of an* achronal *set S consists of all $x \in M$ for which every nbhd U of x contains points y and z and two timelike curves from y to z, of which just one intersects S.*[246]

3. A **wannabe Cauchy surface** *is an* acausal edgeless *(and hence closed) subset of M.*[247]

Wannabe Cauchy surfaces are a "second best" in the absence of Cauchy surfaces (i.e. of global hyperbolicity).[248] A sufficient condition for their existence is the existence of a time function (see §5.9).[249] Simple examples (that are not Cauchy surfaces) are the hyperboloids in $(\mathbb{M}, \eta)$:

$$S = \pm H_1^3; \qquad H_C^\pm(S) = \emptyset; \qquad H_C^\mp(S) = \partial I^\pm(0), \tag{5.181}$$

and the x-axis in the Quinten space-time $(\mathbb{M}_2', \eta_2)$, see §10.7. Cauchy horizons of wannabe Cauchy surface in black hole space-times provide important causal information about their interiors (see chapter 9). In the PDE approach to GR they arise when some MGHD is extendible, see §10.5, and a Cauchy surface for the MGHD turns into a partial one for the extension.

[245] In $2d$ Minkowski space, take $S = [-1,1] \times \{0\}$. Then $D^+(S)$ consists of the triangle with vertices $(-1,0)$, $(1,0)$, and $(0,1)$, whose two upper sides comprise $H_C^+(S)$. Removing $(0,0)$ from S removes the double cone with vertices $(0,0)$, $(-\frac{1}{2},\frac{1}{2})$, $(0,1)$, and $(\frac{1}{2},\frac{1}{2})$ from $D^+(S)$, whereas $H_C^+(S)$ now consists of two zig-zag teeth (draw!).

[246] See Penrose (1972), §5.6; the definition of an edge in Hawking & Ellis (1973), p. 202 or Minguzzi (2019), §2.18 is equivalent *provided S is closed*. By Corollary 2.142 in Minguzzi (2019), a closed acausal or achronal subset $S \subset M$ (think of a spacelike hypersurface) is edgeless (i.e. a wannabe Cauchy surface) iff $S \cup I^+(S) \cup I^-(S)$, the set of all points through which an inextendible timelike curve exists *that intersects* S, is open. For a *Cauchy surface* this set equals M, see (5.183), and is a maximal achronal set (Minguzzi, 2019, Proposition 3.37).

[247] Clearly, $\overline{S} \backslash S \subset \text{edge}(S) \subset \overline{S}$ so that if S is ***edgeless*** in the sense that $\text{edge}(S) = \emptyset$, then S is closed, as claimed. These are usually called ***partial Cauchy surfaces***, but this is bad since some part of a Cauchy surface is not a partial Cauchy surface because it has an edge. Our terminology has the disadvantage that a Cauchy surface is also a wannabe Cauchy surface, but since some wannabes actually make it (e.g. Mick Jagger), this is the lesser evil.

[248] By Theorem 2.146 in Minguzzi (2019), edgelessness is *necessary* for maximality of an achronal set.

[249] See Theorems 3.39 and 4.100 in Minguzzi (2019). This is a far weaker condition than global hyperbolicity!

Without proof we now collect some key properties of all these sets:[250]

Lemma 5.37 *1. Edgeless achronal subsets are closed* topological *hypersurfaces in M.*

2. Future/past Cauchy horizons $H^\pm_C(S)$ of closed sets S are closed and achronal.

3. If S is closed and achronal, then $\partial D^\pm(S) = H^\pm(S) \cup S$.

4. If S is closed and achronal, then $\mathrm{edge}(H^\pm(S)) = \mathrm{edge}(S)$.

5. If S is closed and acausal, then $H^\pm_C(S) \cap S = \mathrm{edge}(S)$.

6. If S is a wannabe Cauchy surface, then $D^\pm(S)\backslash S$ is open and $H^\pm_C(S) \cap S = \emptyset$, so that

$$H^\pm_C(S) = \partial D^\pm(S)\backslash S; \qquad H_C(S) = \partial D(S). \tag{5.182}$$

We also have the following characterization of "true" Cauchy surfaces among the wannabes:

Proposition 5.38 *A wannabe Cauchy surface (or more generally a closed acausal set) $S \subset M$ is a Cauchy surface iff one (and hence all) of the following equivalent conditions are satisfied:*

1. *$D(S) = M$, or equivalently $D^\pm(S) = J^\pm(S)$;*
2. *$H_C(S) = \emptyset$, or equivalently $H^+_C(S) = H^-_C(S) = \emptyset$;*
3. *Every inextendible curve of fixed causality class C intersects S exactly once, where C may (equivalently) be taken to be timelike, causal, or lightlike.*

In particular, a Cauchy surface has empty Cauchy horizon, and yields M as a disjoint union

$$M = S \sqcup I^+(S) \sqcup I^-(S). \tag{5.183}$$

The implication $1 \Rightarrow 2$ is trivial for the first members (which imply the second). Conversely, if $H_C(S) = \emptyset$ then $D(S)$ must be closed, but by Lemma 5.37.6 it is also open. Since M is connected, $D(S) = M$. Furthermore, $1 \Leftrightarrow 3$ (causal case) is almost true by definition of $D(S)$ and of a Cauchy surface. The equivalences within 3 are quite technical, and we omit the proofs.[251]

[250] No. 1 is Lemma 3.17 in Penrose (1972) or O'Neill (1983), Proposition 14.25 and Corollary 14.26. A topological hypersurface is defined as in the second part of Definition 4.13, assuming all maps to be continuous. No. 2 is Proposition 3.15 in Minguzzi (2019). No. 3 is eq. (3.2) in Minguzzi (2019). No. 4 is Proposition 3.22 in Minguzzi (2019). No. 5 is (3.6) in Minguzzi (2019). No. 6 follows from 3 and 5, cf. Corollary 3.26 in Minguzzi (2019).

[251] See O'Neill (1983), Lemma 14.29 for the implication timelike $\Rightarrow$ causal (the converse is trivial). See also Ringström (2009), §10.2.7, who proves 1 from 3 (timelike case), but this includes a proof of timelike $\Rightarrow$ causal in 3. Given $1 \Leftrightarrow 3$ (timelike case), Minguzzi (2019), Theorem 3.40, proves lightlike $\Rightarrow$ timelike (whose converse follows from timelike $\Rightarrow$ causal $\Rightarrow$ lightlike). Note that in Lemma 14.29 O'Neill cannot exclude the case where a causal curve hits S more than once, but in Proposition 5.38 we *assume* S is acausal, which excludes this by definition. This time assuming that S *is* acausal, O'Neill (1983), Corollary 14.54, also directly (i.e. without assuming $1 \Leftrightarrow 3$, timelike case) proves the implication lightlike $\Rightarrow$ timelike. Note that both Hawking & Ellis (1973) and Minguzzi (2019) define Cauchy surfaces as closed *acausal* sets S for which $D(S) = M$, whereas Geroch (1970) and Penrose (1972) define them as *achronal* sets S for which $D(S) = M$, which is equivalent to Definition 5.32. Take $t = 1$ for $x \leq -1$, $t = x$ for $-1 < x < 0$, and $t = 0$ for $x \geq 0$, in $2d$ Minkowski space-time. This achronal set defines a Cauchy surface for Geroch and Penrose but not for Hawking & Ellis and Minguzzi. However, in view of Theorem 5.44 we may always take S to be spacelike, and if we do, then by Lemma 14.42 in O'Neill (1983) it is automatically acausal.

5.9 Time functions

Using a new technique, in this section we finish our sketch of the proof of Theorem 5.34. First, we state a result that is important for both "wannabe" and "genuine" Cauchy surfaces:

Proposition 5.39 *If $S \subset M$ is achronal, the interior* $\mathrm{int}(D(S))$ *of $D(S)$ is globally hyperbolic.*

Here global hyperbolicity is meant as in Definition 5.27. We will prove this proposition shortly, but already note that it implies that a space-time with a Cauchy surface S is globally hyperbolic, cf. Theorem 5.34.3. For given such an S, we have $\mathrm{int}(D(S)) = \mathrm{int}(M) = M$ by Propositions 5.38.1. The converse inference from global hyperbolicity à la Definition 5.27 to a Cauchy surface uses completely different arguments, which will be given later in this section.

We now sketch a proof of Proposition 5.39, based on criterion 2 in Theorem 5.34.[252] If necessary moving the Cauchy surface S in $\mathrm{int}(D(S))$, as in the comments below (5.166), we may place $x \in D^-(S)$ and $y \in D^+(S)$ (this move is not strictly necessary for the argument).[253] There are two ways for the uniform bound (5.167) to fail. One is the possibility of closed causal loops (or more generally imprisoned inextendible curves), but these would cross S many times and this is excluded because S is achronal, cf. Theorem 5.33.1. Secondly, $J(x,y)$ may not be compact–not because it is unbounded but because it is not closed because of missing some points. To understand the link with (5.167), we recall the assumption that the auxiliary Riemannian metric h measuring arc length be *complete* in Theorem 5.34.2. This implies that h-geodesics must avoid such missing points (for otherwise they would be incomplete), and they do so by *increasing* h-arc length near the missing points. Take, for example, Minkowski space-time with the origin removed. The Euclidean metric δ is incomplete on $\mathbb{R}^4\backslash\{0\}$, but the metric

$$h(x) = \delta / \|x\|^2, \tag{5.184}$$

where $\|\cdot\|$ is the Euclidean norm, is complete.[254] Thus the h-arc length of a curve increases arbitrarily as it approaches the origin, and hence only infinitely long curves approach the origin. This behaviour near missing points is generic. Consequently, if the bound (5.167) is violated, there must be causal curves from x to y coming arbitrarily close to missing points in $J(x,y)$ and then by Lemma 5.22 there will also be inextendible causal curves from either x or y to these points, either of which does not cross S, contradicting S being a Cauchy surface in $\mathrm{int}(D(S))$. This argument can be made rigorous by quoting the following generalization of Lemma 5.26:[255]

Lemma 5.40 *Let $(c_n : [0,b_n] \to M)$ be a sequence of fd continuous causal curves from x to $y \neq x$ parametrized by h-arc length, i.e. $c_n(0) = x$ and $c_n(b_n) = y$. There are two possibilities:*

- *Either $b_n \to b < \infty$, in which case there exist a fd continuous causal curve*

$$c : [0,b] \to M, \tag{5.185}$$

and a subsequence of (c_n) that converges to c (in the sense of Definition 5.23);

[252] The "official" proof in Hawking & Ellis (1973), Proposition 6.6.3, or O'Neill (1983), Theorem 14.38, is very hard to understand, though apparently uncontroversial. A much clearer version of it is given by Chruściel (2011), Theorem 2.9.9, but the argument is still very involved. We therefore take a somewhat different route.

[253] One may instead use Lemma 6.6.4 in Hawking & Ellis (1973), whose proof is very clear: if $x \in D^+(S)\backslash H^+_C(S)$, then every past-inextendible causal curve through x intersects $I^-(S)$, and likewise for future–inextendible causal curves, so that every inextendible curve through $x \in \mathrm{int}(D(S))$ intersects both $I^+(S)$ and $I^-(S)$.

[254] Continuing footnote 215: For (5.184), where $\tilde{h} = \delta$, we have $r(x) = \|x\|$ and $\omega(x) = 1/r(x)$ does the job.

[255] See Minguzzi (2019), Theorem 2.53, whose case (ii) was excluded in Lemma 5.26 by global hyperbolicity.

- *Or $b_n \to \infty$, in which case there exist a fd future intextendible continuous causal curve*

$$c : [0,\infty) \to M, \tag{5.186}$$

with $c(0) = x$ and a subsequence of (c_n) that converges (uniformly) to c, as well as a pd past intextendible continuous causal curve $d : (-\infty, 0] \to M$ with $d(0) = y$, and subsequences (c_{n_k}) of (c_n) and (b_{n_k}) of (b_n) such that $d_k(t) := c_{n_k}(t + b_{n_k}) \to d$.

In the second case, the limit curve c starting at x somehow fails to reach y (acquiring infinite h-arc length by wandering around), whereas d, starting at y and "moving back in time", similarly fails to reach x. This is the situation mentioned in the heuristic part of the proof: at least one of these curves fails to reach S and hence S could not be a Cauchy surface in $\mathrm{int}(D(S))$. □

The converse implication from Definition 5.27 to the existence of a Cauchy surface is very different and is based on the construction of a *time function*:

Definition 5.41 *A* **time function** $\mathtt{t} : M \to \mathbb{R}$ *is a continuous surjection that strictly increases along any fd continuous causal curve.*

We now show that global hyperbolicity implies the existence of time functions,[256] having further properties guaranteeing that each level set

$$\Sigma_t := \{x \in M \mid \mathtt{t}(x) = t\} \tag{5.187}$$

is a Cauchy surface. Thus we do not get one Σ but a whole family (Σ_t), which foliates M by

$$M = \sqcup_{t\in\mathbb{R}} \Sigma_t. \tag{5.188}$$

To construct $\mathtt{t}$, we once again take a complete Riemannian metric h on M, as well as some at most countable open cover (V_n) with precompact elements (i.e. $\overline{V_n}$ is compact for each n), so that $M = \cup_n V_n$, with some associated partition of unity (ϕ_n) subordinate to the cover.[257]
We then turn the standard Riemannian measure μ_h induced by h,[258] into a probability measure $\nu_h = \chi\mu_h$, where the function $\chi : M \to \mathbb{R}^+$is defined by $\chi = \sum_n 2^{-n}\phi_n / \int_{V_n} d\mu_n \phi_n$. Without any assumption on a space-time (M,g), this measure is: i) *finite* (i.e. $\nu_h(A) < \infty$ for any Borel measurable $A \subset M$); ii) *open*, in that $\nu_h(U) > 0$ for any open set $U \subset M$; iii) *regular* (in the usual sense of measure theory);[259] and iv) assigns zero measure to the achronal boundaries $\partial I^{\pm}(x)$.

Any measure with these properties can be used in the following construction. Define

$$V^{\pm} : M \to \mathbb{R}^+; \qquad \mathtt{t} : M \to \mathbb{R}; \tag{5.189}$$

$$V^{\pm}(x) := \nu_h(J^{\pm}(x)); \qquad \mathtt{t}(x) := \ln\left(\frac{V^-(x)}{V^+(x)}\right). \tag{5.190}$$

[256] The existence of a time function is equivalent to the weaker assumption of ***stable causality***, which means that a space-time (M,g) has a Lorentzian metric g' such that (M,g') is causal and $g(X,X) \leq 0$ implies $g'(X,X) < 0$. See e.g. Minguzzi & Sánchez (2008), Definition 3.52 and Theorem 3.56. Global hyperbolicity yields (5.191) below.

[257] This means that $\varphi_n \in C_c^\infty(V_n)$ and $\sum_n \varphi_n(x) = 1$ for all $x \in M$.

[258] This measure is defined intrinsically, but in coordinates we have $d\mu_h(x) = \sqrt{\det h(x)}\, dx^0 \cdots dx^3$. See §7.1.

[259] This means that for any Borel set $A \subset M$ one has outer regularity $\nu_h(A) = \inf\{\nu_h(U) \mid U \supset A, U \subset M \text{ open}\}$ as well as inner regularity $\nu_h(A) = \sup\{\nu_h(K) \mid K \subset A, K \subset A, K \text{ compact}\}$. This follows from the fact that ν_h is equivalent to Lebesque measure in any local chart, which also implies the last property, given that the achronal boundaries $\partial I^{\pm}(x)$ have dimension 3 and ν_h is supported in dimension 4.

Now very simple examples (e.g. Quinten space-time) show that V^- and V^- may easily be discontinuous, but compactness of all double cones is sufficient to make both functions continuous; in fact, for any sequence $x_n \to x$ one then has $1_{J^\pm(x_n)} \to 1_{J^\pm(x)}$ a.e.[260]

Furthermore, any of the assumptions of causality, non-imprisonment, or strong causality (which are equivalent when all double cones are compact, see below) suffice to prove that:[261]

1. V^- strictly increases and V^+ strictly decreases along fd causal curves. This is natural, as moving forward in time leaves more causal past behind and anticipates less causal future.

2. Along any inextendible causal curve $c : \mathbb{R} \to M$ (parametrized by h-arc length) one has

$$\lim_{t\to\infty} V^+(c(t)) = \lim_{t\to-\infty} V^-(c(t)) = 0; \tag{5.191}$$

$$\lim_{t\to\pm\infty} \mathtt{t}(c(t)) = \pm\infty. \tag{5.192}$$

See Lemma 5.22 for the domain $\mathbb{R}$. Eq. (5.191), which implies (5.192), is also quite intuitive, for if there had been any causal future $J^+(x)$ left beyond the end of the curve, then $c(\cdot)$ could have been extended into it and hence would not have been future inextendible. Similarly, no causal past is left before the beginning of the curve.[262] This implies that $\mathtt{t}$ as defined in (5.190) has the right properties to serve as a time function, which strictly monotonically increases from $-\infty$ to $+\infty$ along any fd causal curve parametrized by h-arc length. This also means that any such curve hits each set Σ_t as defined by (5.187) once, which makes all Σ_t Cauchy surfaces. □

This finishes the sketch of the proof of Theorem 5.34, but there is much more to say about the potential smoothness of the Cauchy surfaces Σ_t as well as of the time function $\mathtt{t}$. We first sharpen the definition of a time function (see Definition 5.41 in the following way:

Definition 5.42 *A* **temporal function** *is a smooth surjection* $\mathtt{t} : M \to \mathbb{R}$ *with timelike past-directed gradient* $\nabla\mathtt{t} = \sharp(d\mathtt{t})$*, or, in coordinates,* $\nabla^\mu \mathtt{t} = g^{\mu\nu}\partial_\nu \mathtt{t}$.

Temporal functions are time functions, for if $c : I \to M$ is fd timelike, then $\dot{c}(\mathtt{t}) = g(\nabla\mathtt{t}, \dot{c})$ is strictly positive along $c(\cdot)$.[263] If $\mathtt{t}$ is a temporal function, then the level set Σ_t as defined in (5.187) is spacelike, since for any $x \in \Sigma_t$ and $X \in T_xM$ we have $g_x(\nabla\mathtt{t}, X) = X\mathtt{t}(x)$, which by (5.187) vanishes for any $X \in T_x\Sigma_t$. This forces $g_x(X,X) > 0$ by the following lemma.

Lemma 5.43 *For any Lorentzian metric* g*, if* $g(T,T) < 0$ *and* $g(T,X) = 0$*, then* $g(X,X) > 0$.

Proof. Taking an orthonormal basis reduces this to the Minkowski case. Let $T = (T_0, \vec{T})$ and $X = (X_0, \vec{X})$. Then $T_0^2 > \|\vec{T}\|^2$ and $T_0X_0 = \vec{T}\cdot\vec{X}$, so that $|T_0X_0| \leq \|\vec{T}\|\|\vec{X}\|$. This implies $X_0^2 < \|\vec{X}\|^2$, since $X_0^2 \geq \|\vec{X}\|^2$ gives a contradiction. For example, if $T_0 > 0$ and $X_0 \geq 0$ the assumptions give $T_0 > \|\vec{T}\|$ and $T_0X_0 \leq \|\vec{T}\|\|\vec{X}\|$, which contradict $X_0 \geq \|\vec{X}\|$. The other three cases (i.e. $T_0 > 0$ and $X_0 \leq 0$, $T_0 < 0$ and $X_0 \geq 0$, and $T_0 < 0$ and $X_0 \leq 0$) are similar. □

[260] See e.g. Chruściel (2011), §2.11 for very clear proofs of this and the following properties.

[261] The stronger property of global hyperbolicity is needed to prove (5.191), which implies (5.192). See Minguzzi & Sánchez (2008) as well as the papers by the latter and Bernal cited in footnote 238.

[262] To illustrate what happens at a more technical level, let us show that $x \mapsto \nu_h(V^-(x))$ is strictly increasing along fd causal curves c. Take x and $y \in J^+(x)$ on c; then $y \notin J^-(x)$ by causality. Global hyperbolicity also guarantees that $J^-(x)$ is closed, so that its complement in M is open and hence y has an open nbhd U disjoint from $J^-(x)$. But $J^-(x) \subset J^-(y)$ and $\nu_h(U \cap J^-(y)) = \nu_h(U \cap I^-(y)) > 0$, by the properties of ν_h, so that $\nu_h(J^-(y)) > \nu_h(J^-(x))$.

[263] Here $\dot{c}(\mathtt{t}) : C(I) \to M$ applies the tangent vector $\dot{c}$ to the function $\mathtt{t}$, not te be confused with $\dot{c}_{c(t)} \in T_{c(t)}M$.

We then have the following result, which smoothens out earlier topological properties:

Theorem 5.44 *A space-time (M,g) is globally hyperbolic iff it has a smooth spacelike Cauchy surface. In that case:*

1. *There exists a smooth temporal function $\mathrm{t}: M \to \mathbb{R}$ such that M is foliated as in (5.188), where each Σ_t is a smooth spacelike Cauchy surface and all Σ_t are diffeomorphic.*
2. *M is diffeomorphic to $\mathbb{R}\times\Sigma$, where Σ is diffeomorphic to Σ_t for any $t\in\mathbb{R}$.*
3. *(M,g) is isometric to $(\mathbb{R}\times\Sigma, g')$, where the metric g' is given in the $3+1$ form*

$$g' = -L^2 dt^2 + \tilde{g}, \tag{5.193}$$

in which $L:\mathbb{R}\times\Sigma\to(0,\infty)$ is the (smooth) **lapse** *function and $\tilde{g}$ is a (possibly time-dependent) Riemannian metric on the Cauchy surface Σ.*

This landmark theorem is due Bernal and Sánchez.[264] The proof, which is very technical, constructs a temporal function t, which gives a time orientation on M via the vector field

$$T = -\nabla \mathrm{t}. \tag{5.194}$$

In Minkowski space-time, with $\mathrm{t} = x^0$, this would be $T = \partial_t$, whence the minus sign in T. This implies claim 2, for in the construction of the map $\mathbb{R}\times\Sigma\to M$ sketched after (5.166) one can take T proportional to $\nabla\mathrm{t}$. The remainder of the proof requires machinery beyond our scope.

Definition 5.27 and Theorem 5.34 state the various definitions of global hyperbolicity as they have been used for about the first 50 years of Lorentzian causality theory, except that in Definition 5.27 strong causality has traditionally been used instead of non-imprisonment (and even the still weaker causality property could have been used).[265] However, given compactness of the double cones, if $\dim(M)\geq 3$ it turns out to be sufficient for global hyperbolicity to require the extremely weak condition that (M,g) be ***non-totally vicious***, which means that there need just be a single point through which no closed timelike curve passes.[266] If all $J^{\pm}(x)$ are closed (which, as Lemma 5.29 shows, is a consequence of compactness of the double cones), this implies that (M,g) is strongly causal (and hence non-imprisoning). Moreover, if (M,g) is totally vicious (i.e. there is a closed causal curve through each point), then $J^{\pm}(x) = M$ for each x. If it is also required that all $J(x,y)$ are compact, this forces M to be compact. Hence if M is non-compact and all double cones are compact, then (M,g) cannot be totally vicious. In conclusion, under physically reasonable assumptions global hyperbolicity has reached a very simple form:[267]

Proposition 5.45 *Let (M,g) be a space-time with $\dim(M)\geq 3$ and M non-compact. Then (M,g) is globally hyperbolic iff all double cones $J^+(x)\cap J^-(y)$ are compact.*

[264] See references in footnote 238. Apart from these, see also Ringström (2009), chapter 11. Other constructions of temporal functions were given by Fathi & Siconolfi (2012) and Chruściel, Grant, & Minguzzi (2016).

[265] See Bernal & Sánchez (2006) and Minguzzi (2019).

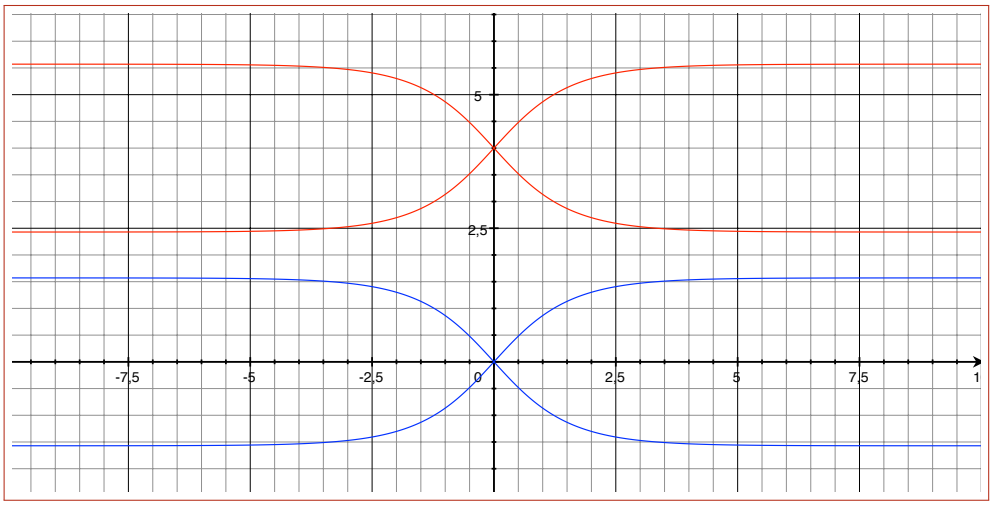
[266] And hence a space-time is ***totally vicious*** if some closed timelike curve passes through every point.

[267] For a complete proof see Hounnonkpe & Minguzzi (2019), which relies on results by Clarke & Joshi (1988). Briefly, the latter proved the inference from non-total viciousness to chronology (assuming an even weaker property than closedness of all sets $J^{\pm}(x)$, namely that (M,g) is ***reflecting***, which means that $x\in\overline{J^-(y)}$ iff $y\in\overline{J^+(x)}$), which the former then strengthened to strong causality. This simplification is quite remarkable, after 50 years!

5.10 Global hyperbolicity: AdS as a counterexample

After all this abstraction, some specific examples may be welcome. Minkowski space-time $(\mathbb{M},\eta)$ is globally hyperbolic as well as geodesically complete, and the removal of any point from it ruins both properties in a somewhat trivial way. Much more interesting examples arise from the two combinations *geodesically complete but non-globally hyperbolic* and the opposite, i.e. *geodesically incomplete but globally hyperbolic*. The latter will be the subject of chapters 6 and 9. A nice example of the former is *anti de Sitter space-time* AdS^n_ρ, defined for any $n = \dim(AdS^n_\rho) \geq 2$ and $\rho > 0$, see §4.4. To make our point, the case $n = 2$, $\rho = 1$ suffices, for which we simply write $AdS \equiv AdS^2_1$; the conclusions will be true for any $n \geq 2$ and $\rho > 0$.

Eq. (4.93) gives AdS as the set of all $(x_{-1},x_0,x_1) \in \mathbb{R}^3$ such that $x^2_{-1} + x^2_0 = x^2_1 + 1$, with metric induced from $\eta' = \mathrm{diag}(-1,-1,+1)$. This space is homeomorphic to $S^1 \times \mathbb{R}$ and contains closed timelike curves, such as $t \mapsto (\cos t, \sin t, 0)$ defined on $t \in [0, 2\pi]$. For this reason alone it cannot be globally hyperbolic, but its Lorentzian cover $\widetilde{AdS}$, where these curves are defined for all $t \in \mathbb{R}$ and no longer close, isn't either, as it fails on compactness of double cones $J(x,y)$.



*Some lightlike geodesics in 2d anti de Sitter space-time. The χ-axis is horizontal, the τ-axis is vertical. The blue lines are the two lightlike geodesics through the origin $x = (0,0)$, cf. the corresponding cross in 2d Minkowski space-time $\mathbb{M}_2$. These blue curves asymptote to $\pm\frac{1}{2}\pi$ as $|x| \to \infty$. Similarly, the red lines are the lightlike geodesics through $y = (0,4)$. The set $J^+(x)$ is the area above the two upper blue lines (including boundary) whilst $J^-(y)$ is the area below the two lower red lines (*idem*). The double cone $J^+(x) \cap J^-(y)$ stretches on forever to the left and to the right and hence is not compact, violating global hyperbolicity.*

To see this,[268] introduce local coordinates $(\tau,\chi) \in (-\pi,\pi) \times \mathbb{R}$ initially on AdS by

$$x_{-1} = \cos\tau\cosh\chi; \qquad x_0 = \sin\tau\cosh\chi; \qquad x_1 = \sinh\chi. \tag{5.195}$$

In these coordinates, the metric on AdS is simply given by

$$ds^2 = -\cosh^2\chi\, d\tau^2 + d\chi^2, \tag{5.196}$$

and therefore $\widetilde{AdS}$ can be globally coordinatized by $(\tau,\chi) \in \mathbb{R}^2$, with the same metric (5.196).

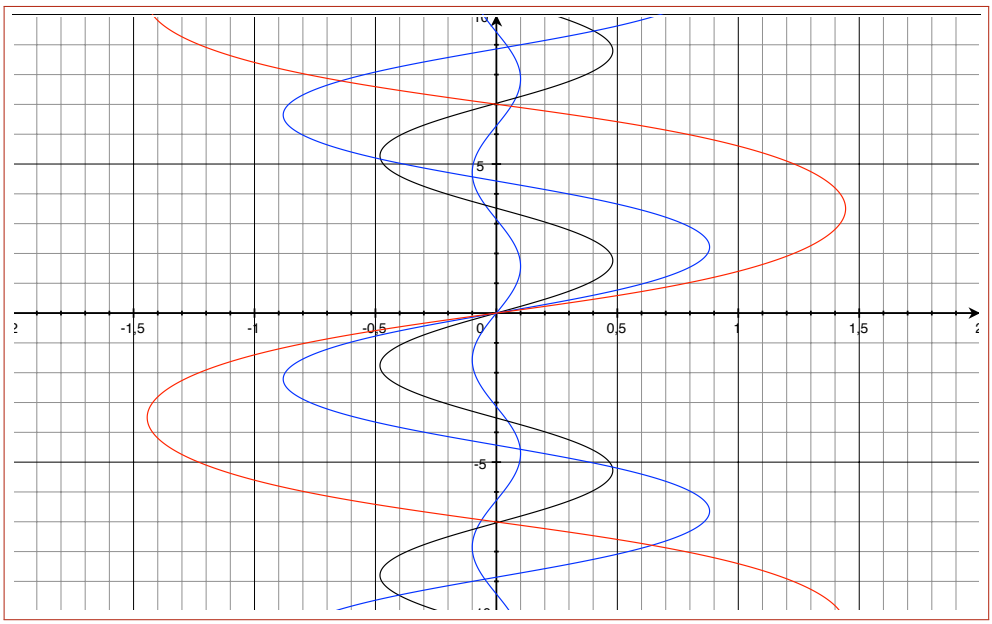
[268] Let $\mathbb{R}^3$ have any (semi) Riemannian metric g'. Take a surface $\Sigma = F(U) \subset \mathbb{R}^3$ defined by a smooth injective function $F : U \to \mathbb{R}^3$ satisfying the conditions stated at the beginning of §4.3, where $U \subset \mathbb{R}^2$ is open. Then the induced metric g on Σ is given by $g_{\mu\nu}(u^1,u^2) = \sum_{i,j=1,2,3} g'_{ij} \frac{\partial F^i(u^1,u^2)}{\partial u^\mu} \frac{\partial F^j(u^1,u^2)}{\partial u^\nu}$, where $\mu,\nu = 1,2$.

Taking $\chi(t) = t$, lightlike (pre)geodesics in $\widetilde{AdS}$ are solutions of $\dot{\tau}_\pm(t) = \pm 1/\cosh t$. Thus

$$\tau_\pm(\chi) = \pm 2 \arctan(\tanh(\tfrac{1}{2}\chi)), \tag{5.197}$$

gives the lightlike (pre) geodesics through $(0,0)$ are whilst those through $(0,4)$ are the same, moved up by $(0,4)$. In the (τ,χ) plane these give the blue and red curves, respectively.

Timelike geodesics in *AdS* are also quite remarkable, as the following picture shows.



Some timelike geodesics γ_C through the origin in AdS. The inner blue one has $C = 0.1$, the black one has $C = 0.5$, then next blue one has $C = 1$, and the outer red curve has $C = 2$. For $C = 0$ the geodesic is simply the τ-axis; the geodesics corresponding to negative values of c are the mirror images (in the τ-axis) of those displayed. All geodesics spiral around the τ-axis and continue to focus and defocus. Mind the difference in scale between the axes!

The geodesic equations for the metric (5.196) are easily found to be

$$\ddot{\tau} + 2\tanh\chi \cdot \dot{\tau}\dot{\chi} = 0; \tag{5.198}$$

$$\ddot{\chi} + \cosh\chi \sinh\chi \cdot \dot{\tau}^2 = 0, \tag{5.199}$$

and one can explicitly find all timelike geodesics through the origin, namely

$$\chi_C(t) = \operatorname{arcsinh}(C \cdot \sin t); \tag{5.200}$$

$$\tau_C(t) = \sqrt{1+c^2} \int_0^t ds\,(1 + C^2 \sin^2 s)^{-1}, \tag{5.201}$$

where $C \in \mathbb{R}$ is a constant,[269] physically interpreted via $\dot{\chi}(0) = C$ and $\dot{\tau}(0) = \sqrt{1+C^2}$. These geodesics $\gamma_C(t) = (\chi_C(t), \tau_C(t))$ are all timelike, with $g(\dot{\gamma}_C(t), \dot{\gamma}_C(t)) = -1$ for all C and $t \in \mathbb{R}$.

Combining the two plots gives another way to see that *AdS* is not globally hyperbolic, because vast areas in $J^+(0,0)$ are inaccessible by timelike curves from the origin, eternally attracted to the τ-axis as these apparently are. Thus global hyperbolicity would contradict Theorem 5.30.

[269] For $0 < t < \frac{1}{2}\pi$ one has $\tau(t) = \arctan(\sqrt{1+c^2} \cdot \tan t)$, but the formula in the main text is more useful.

6 The singularity theorems of Hawking and Penrose

We now turn to the famous singularity theorems of Hawking and Penrose.[270] The possibility of singular space-times in GR was suggested by three of the earliest exact solutions to Einstein's equations, namely the ***Schwarzschild solution*** from 1916, de ***de Sitter universe*** from 1917 (here given in Eddington's coordinates from 1922, see §9.1), and the ***Friedman–Lemaître–Robertson–Walker (FLRW) solution*** first found (by Friedman) in 1922. These are given by

$$ds_S^2 = -\left(1-\frac{2m}{r}\right)dt^2 + \left(1-\frac{2m}{r}\right)^{-1} dr^2 + r^2 d\Omega; \tag{6.1}$$

$$ds_{dS}^2 = -\left(1-\frac{r^2}{\rho^2}\right)dt^2 + \left(1-\frac{r^2}{\rho^2}\right)^{-1} dr^2 + r^2 d\Omega; \tag{6.2}$$

$$ds_F^2 = -dt^2 + a(t)^2(d\chi^2 + f(\chi)^2 d\Omega), \tag{6.3}$$

respectively, where ds^2 is just the original notation for the metric, and $d\Omega$ is defined in (4.72).

- In the vacuum Schwarzschild solution (6.1), $m > 0$ is the mass of some gravitating object and the space-time is $M_S = \mathbb{R} \times \Sigma$, where at least initially, in polar coordinates (r, θ, φ), the spatial part $\Sigma \subset \mathbb{R}^3$ is restricted to $r > 2m$. Here the value $r = 2m$ looks threatening, as does $r = 0$ (although the latter is not, as yet, in the domain of the solution).[271]

- The de Sitter metric (6.2), initially defined as above but now for $0 \leq r < \rho$, requires a cosmological constant $\lambda = 3/\rho^2$, where ρ is the radius of the visible universe. The potential danger lies at $r = \rho$, which looks as bad as $r = 2m$ in the Schwarzschild metric.

- The FLRW solution (6.3) requires matter. The space-time is $M = (0, \infty) \times \Sigma$, where:

  - $\Sigma = S^3$ (the 3-sphere) and $f(\chi) = \sin\chi$ for $k = 1$ (positive curvature);
  - $\Sigma = \mathbb{R}^3$ and $f(\chi) = \chi$ for $k = 0$ (zero curvature),
  - $\Sigma = H^3$ (the $3d$ hyperboloid) and $f(\chi) = \sinh\chi$ for $k = -1$ (negative curvature).

  The function $a(t)$ depends on the precise matter content of the universe. For example, for a dust-filled spatially flat universe one has $a(t) \sim t^{2/3}$ as $t \to 0$, where also the Ricci scalar R blows up. The precise form of $R(t)$ again depends on the matter content, but in the same dust-filled case one finds $R(t) \sim t^{-2}$. Similarly for other forms of matter. See also §8.3.

Even Hilbert and Einstein were initially confused about the meaning of these apparent or real singularities, but today it is clear that $r = 2m$ and $r = \rho$ are just singularities of the coordinate systems in which the Schwarzschild and de Sitter solution are expressed. This is not to say that nothing interesting happens at $r = 2m$: The hypersurface $r = 2m$ is an *event horizon*, see chapters 9 and 10, and even the utterly regular (even constant-curvature) de Sitter space-time has a kind of horizon at $r = \rho$, see §9.1. On the other hand, in the Schwarzschild solution $r = 0$, if it were part of space-time, would be a real singularity of the metric and its associated tensors.[272]

[270] Israel (1987), Tipler, Clarke & Ellis (1980), Clarke (1993), Earman (1995, 1999), Earman & Eisenstaedt (1999), Senovilla (1998), Joshi (2014), Senovilla & Garfinkle (2015), and Curiel (2019) survey singularities and singularity theorems in GR, including history. The classical textbook exposition remains Hawking & Ellis (1973).

[271] We will study this solution in detail in §9.2. Mathematically, it also makes sense for $m < 0$, see §9.5.

[272] The singularity is detected by the Kretschmann scalar $R^{\rho\sigma\mu\nu}R_{\rho\sigma\mu\nu}$, which goes like r^{-6} as $r \to 0$, see (9.18).

In his (very brief) published reply to de Sitter and his universe,[273] Einstein in some sense paved the way to the modern notion of a singularity in arguing about de Sitter's metric (6.2) that:

1. The singularity in the metric [at $r = \rho$] is real since 'it seems that no choice of coordinates can remove this discontinuity' [immaterially, Einstein said this in different coordinates].
2. "Singular' points [at $r = \rho$] can be reached from "regular" points in finite proper time.
3. The conjunction of 1 and 2 is a 'grave argument against the admissibility of this solution', because this conjunction makes $r = \rho$ 'a genuine singularity' (*eine echte Singularität*').

Although claim 1 is simply wrong (de Sitter space is regular in every conceivable way), and criterion 2 is stated incorrectly (one should use timelike *geodesics* instead of the arbitrary timelike *curves* Einstein mentions in his paper),[274] the underlying idea was surely forward-looking![275]

Talk about singularities in GR remained ambiguous until the 1960s. In 1963, Misner first argues that everything should be described in terms of a regular manifold M carrying a regular metric g, so that singularities cannot be "in" space-time (this marks a decisive difference with say singularities in for example the electro-magnetic field).[276] Attempting to capture what it could mean for a Lorentzian manifold (M, g) to be "singular", Misner then requires the implications

$$\textit{curvature singularity} \;\Rightarrow\; (M,g)\textit{ is "singular"} \;\Rightarrow\; (M,g)\textit{ is geodesically incomplete} \quad (*)$$

where 'curvature singularity' means unbounded curvature along a (semi) open geodesic segment, cf. Proposition 6.2 below. He adds that it is 'commonly accepted' that an 'essentially singular space' is not only incomplete, but also inextendible (cf. Definition 6.1). Hence we have both necessary and sufficient conditions for a space-time to be singular, but not yet a definition. Inspired by Penrose's paper from 1965 discussed below, it may have been Hawking (1966), §6.1, i.e. in his Adams Prize Essay, who, explicitly assuming inextendibility, first proposed to:[277]

> take timelike and lightlike geodesic incompleteness as our definition of a singularity of spacetime.

In defense of this definition, Hawking and Ellis make both a physical and a pragmatic point:

> 'Timelike geodesic completeness has an immediate physical significance in that it presents the possibility that there could be freely moving observers or particles whose histories did not exist after (or before) a finite interval of proper time. This would appear to be an even more objectionable feature than infinite curvature and so it seems appropriate to regard such a space as singular. (...) The advantage of taking timelike and/or null incompleteness as being indicative of the presence of a singularity is [also] that on this basis one can establish a number of theorems about their occurrence.' (Hawking & Ellis, 1973, p. 258)

Although this has become the "received definition", one may side with Geroch (1968), p. 526:

(a) there is no widely accepted definition of a singularity in general relativity;

(b) each of the proposed definitions is subject to some inadequacy.

---

[273] This reply is Einstein (1918d), which is Doc. 5 in Einstein (2002a), translated in Einstein (2002b), Doc. 5.

[274] For any timelike curve reaching some point in infinite proper time one can find another timelike curve doing so in finite time, which makes Einstein's way of stating point 2 empty (Clarke, 2003, pp. 2–3).

[275] Einstein (1939) still used this kind of reasoning to explain why the Schwarzschild radius $r = 2m$ was (allegedly) *in*accessible and hence should *not* be seen as a singularity. Lack of stubbornness was not his trait!

[276] See Misner (1963), who acknowledges L. Shapley and L. Marcus as sources of his discussion.

[277] In his PhD Thesis, Hawking (1965), §4.1, says, without any further ado (or mention of inextendibility): 'any model must have a singularity, that is, it cannot be a geodesically complete C^1, piecewise C^2 manifold.'

Remarkably, in the single most famous paper on singularities in GR, Penrose (1965) does not explicitly define what he means by a "singularity", although all uses of the word refer to examples of $r = 0$ curvature singularities, and the general context of stellar collapse also suggests that this is what he means. Nonetheless, what he proves is geodesic incompleteness (see Theorem 6.15), and since Misner's implications (*) do not work the other way round, this leaves a gap.[278] This gap did not reduce the huge and immediate impact of the paper (later Penrose won half of the 2020 Physics Nobel Prize on this basis). Namely, until 1965 it had been quite unclear whether singularities (however defined) were *generic* or *exceptional* (i.e. typical of very special solutions with a high degree of symmetry, and hence absent in realistic solutions).[279] Penrose's work, almost instantly followed by Hawking's, settled this in favour of genericity. Hence despite Geroch's warning,[280] we now formalize the "received definition" (cf. Definition 5.18).

Definition 6.1 *1. A space-time (M,g) (cf. Definition 5.3) is* **extendible** *if there exists a space-time (M',g'), where $\dim(M') = \dim(M)$ and $M' \neq M$, and an isometric embedding $i : M \hookrightarrow M'$ (i.e. $i^*g' = g$) with $i(M)$ open. It is* **inextendible** *if this is not the case.*[281]

2. A space-time (M,g) is **causally incomplete** *if M contains an incomplete causal geodesic.*

3. A space-time (M,g) is **singular** *if it is both causally incomplete and inextendible in a way relevant to its incompleteness, that is, if M contains an incomplete causal geodesic γ such that there is no extension $i : M \to M'$ for which the curve $i \circ \gamma$ is extendible.*[282]

A very useful criterion for proving inextendibility of space-times (M,g) is the following:[283]

Proposition 6.2 *If either all causal (or even just all timelike) geodesics in M are complete, or for any incomplete causal (*ibid.*) geodesic $\gamma : [0,b) \to M$ in M there is a curvature invariant (such as R or $R^{\rho\sigma\mu\nu}R_{\rho\sigma\mu\nu}$, etc.) that is unbounded as $t \to b$, then (M,g) is inextendible.*

[278] It is easy to get confused by what Penrose writes: 'If, as seems justifiable, actual physical singularities in space-time are not to be permitted to occur, the conclusion would appear inescapable that inside such a collapsing object at least one of the following holds: (...) (c) The space-time manifold is incomplete (...)'. A footnote at this point adds that 'The "I'm all right, Jack" philosophy with regard to the singularities would be included under this heading!'. This suggests that he considers incompleteness a logically possible but cheap way to avoid singularities. Just before stating his theorem, he adds that 'the existence of a singularity can never be inferred, however, without an assumption such as completeness of the manifold under consideration'. (Penrose, 1965, p. 58). This looks like the very opposite of defining singularities through geodesic incompleteness! But it all makes sense if "complete" is taken to mean *inextendible*, as both Erik Curiel and José Senovilla propose (in e-mails, May 2021). See also §10.4.

[279] For example, Einstein and Landau's school maintained the latter. Earlier work by Raychaudhuri, Komar, Szekeres, Misner, and Shepley towards the singularity theorems is discusses in the references in footnote 270.

[280] In defense, is is hard to make rigorous the idea that in space-time singularities the curvature blows up (Clarke, 1993; Senovilla, 1998). For example, on the one hand the components of the Riemann and Ricci tensors are coordinate-dependent but on the other hand curvature scalars like R, $R^{\mu\nu}R_{\mu\nu}$, or $R^{\rho\sigma\mu\nu}R_{\rho\sigma\mu\nu}$ do not capture all about curvature. Furthermore, since singularities are not actual points of space-time one cannot speak of nbhds of such points either. There are also space-times that both common sense and Definition 6.1 deem singular although the curvature is bounded. Two examples are the $2d$ Misner space-time (Misner, 1967; Thorne, 1993; Hawking & Ellis, 1973, §5.8), and the flat $4d$ space-time $\mathbb{R}^4/\sim$ whose spatial polar coordinates (r,θ,ϕ) are defined just for $0 < \phi < a$, where $\pi < a < 2\pi$, and $(t,r,\theta,\phi = 0) \sim (t,r,\theta,\phi = a)$. This has a conical singularity at $r = 0$.

[281] This tacitly assumes smoothness of (M',g') and i. One may also keep M' and i smooth but lower the regularity of g'. This is important e.g. for cosmic censorship, see §10.5. See also Senovilla (1998), Definition 3.1.

[282] Adapted from Clarke (1993), p. 10, who uses curves instead of geodesics (cf. §10.4). See also Manchak (2014).

[283] See Remark 5.45 on page 155 of O'Neill (1983) or Proposition 4.4.3 in Chruściel (2020). The proof is by contradiction: if M were extendible, then some such γ could be continued past b into M' with a finite limit as $t \to b$.

6.1 Congruences of geodesics

For the singularity theorems one needs a variation on conjugate points called *focal points*, which are like conjugate points but now defined relative to a *congruence of geodesics* rather than a single one. In general, a ***congruence of curves*** through an open set $U \subset M$ is simply a family of curves such that each point of U lies on exactly one such curve. This is automatic if the congruence arises as the flow of a vector field defined on U, and *vice versa*, a congruence yields a vector field as its tangent, so that congruences in U and vector fields in U are interchangeable. The following constructions on a Lorentzian manifold may be performed in either the timelike or lightlike case. We start with the former; see §6.3 for the latter. Thus we start from a fd timelike vector field $u \in \mathfrak{X}(U)$ defined locally on some open $U \subset M$, normalized such that, at each $x \in U$,

$$g_x(u_x, u_x) = u_\mu(x)u^\mu(x) = -1. \tag{6.4}$$

The associated congruence, then, is obtained by integrating this vector field. In one example of interest, u is the 4-velocity of some (relativistic) fluid moving in the cosmos, but for Hawking's singularity theorem one starts from a spacelike hypersurface $\Sigma \subset M$ (which will eventually, but not yet, be taken to be a Cauchy surface), and defines the congruence as consisting of all timelike geodesics γ emanating from Σ with initial velocities normal to Σ, for as long as they do not cross and do remain timelike. This condition defines the open set $U \supset \Sigma$; singularities arise (outside U) if the geodesics either cross or become lightlike. The right parametrization of each γ then enforces (6.4), where $u = \dot\gamma$; recall that $g(\dot\gamma, \dot\gamma)$ is constant along γ, so that (6.4) persists in t.

The flow φ_t of u (which at $x \in \Sigma$ of course is given by $\varphi_t(x) = \gamma(t)$, where γ is the geodesic emanating from x normal to Σ and satisfying $g(\dot\gamma, \dot\gamma) = -1$), then defines $\Sigma_t = \varphi_t(\Sigma)$, which is diffeomorphic to Σ as long as $\varphi_t(x) \in U$ for all $x \in \Sigma$, and hence one obtains a disjoint family of spacelike hypersurfaces $(\Sigma_t)_{t\in I}$, where $I \subset \mathbb{R}$ is some open interval.[284] Alternatively, a temporal function $\mathsf{t} : U \to \mathbb{R}$ (cf. Definition 5.42) defines a family of spacelike hypersurfaces

$$\Sigma_t = \{x \in U \mid \mathsf{t}(x) = t\}, \tag{6.5}$$

to each of which u remains orthogonal. Indeed, if X is a vector field on Σ, i.e. $X \in T_x\Sigma$, then $\varphi_t'(X) \in T_{\varphi_t(x)}\Sigma_t$ extends X to U. Still calling this extension X, we have $\mathscr{L}_u X = 0$ throughout U and $g_x(u, X) = 0$ for $x \in \Sigma$ by construction. Then

$$\frac{d}{dt}g(u,X) = \nabla_u g(u,X) = g(\nabla_u u, X) + g(u, \nabla_u X) = g(u, \nabla_X u) = \tfrac{1}{2}\nabla_X g(u,u) = 0, \tag{6.6}$$

where we used $\nabla_u u = 0$ (since each γ is a geodesic), torsion-freeness (3.47) of ∇, which gives

$$\nabla_u X = \nabla_X u + [u, X] = \nabla_X u + \mathscr{L}_u X = \nabla_X u, \tag{6.7}$$

and finally (6.4). In terms of t, the unit normal u to each Σ_t is then given by

$$u = -L\nabla\mathsf{t}; \tag{6.8}$$

$$L = 1/\sqrt{-g(\nabla\mathsf{t}, \nabla\mathsf{t})}, \tag{6.9}$$

where the function $L : U \to \mathbb{R}^+_*$ is called the ***lapse***; it will be taken up in §8.1.

[284] This construction fails as soon as some $\gamma(t) = \varphi_t(x)$ reaches a focal point, as discussed in the next section. In that case, the map $\varphi(x, Y) = \exp_x(Y)$ from $N^\Sigma M$ to M (where $N^\Sigma M$ is the *normal vector bundle* over Σ to the embedding $\Sigma \hookrightarrow M$, i.e. $X \in N^\Sigma_x M$ if $X \in T_x M$ and $X \perp T_x\Sigma$, where $x \in \Sigma$), which is a diffeomorphism from a nbhd of the zero section in $N^\Sigma M$ to a tubular nbhd of Σ in M, fails to be a diffeomorphism. Note that $\varphi(x, tu) = \varphi_t(x)$.

A timelike vector field u defines a fair amount of derived tensors, each of some importance:

$$a^\mu := u^\nu \nabla_\nu u^\mu \qquad \textit{(acceleration)}; \tag{6.10}$$
$$h_{\mu\nu} := g_{\mu\nu} + u_\mu u_\nu \qquad \textit{(spatial projection)}; \tag{6.11}$$
$$k_{\mu\nu} := h^\rho_\mu h^\sigma_\nu \nabla_\rho u_\sigma \qquad \textit{(minus extrinsic curvature)}; \tag{6.12}$$
$$\omega_{\mu\nu} := k_{[\mu\nu]} \qquad \textit{(vorticity)}; \tag{6.13}$$
$$\sigma_{\mu\nu} := k_{(\mu\nu)} - \tfrac{1}{3}\theta h_{\mu\nu} \qquad \textit{(shear)}; \tag{6.14}$$
$$\theta := g^{\mu\nu}k_{\mu\nu} = h^{\mu\nu}k_{\mu\nu} \equiv \mathrm{tr}(k) \qquad \textit{(expansion)}, \tag{6.15}$$

where $k_{(\mu\nu)} := \frac{1}{2}(k_{\mu\nu} + k_{\nu\mu})$, $k_{[\mu\nu]} := \frac{1}{2}(k_{\mu\nu} - k_{\nu\mu})$, and $h^\mu_\nu := \delta^\mu_\nu + u^\mu u_\nu$. It follows that

$$k_{\mu\nu} = \tfrac{1}{3}\theta h_{\mu\nu} + \sigma_{\mu\nu} + \omega_{\mu\nu}; \tag{6.16}$$
$$\nabla_\mu u_\nu = k_{\mu\nu} - u_\mu a_\nu; \tag{6.17}$$
$$\theta = \nabla_\mu u^\mu. \tag{6.18}$$

Eq. (6.16) is trivial. Eq. (6.17) can be checked by contracting both sides first with u^μ, then with u^ν, and finally with vectors orthogonal to u. The first contraction merely reproduces the definition (6.10). For the second contraction we use (6.4), (3.65), and (3.52) to compute

$$0 = \partial_\mu g(u,u) = (\nabla_\mu g)(u,u) + g(\nabla_\mu u, u) + g(u, \nabla_\mu u) = 0 + 2g(u, \nabla_\mu u), \tag{6.19}$$

whence $u^\nu \nabla_\mu u_\nu = g(u, \nabla_\mu u) = 0$. The third contraction reproduces the definition (6.12). Eq. (6.18) follows from (6.17) and (6.15), since again $g(a,u) = 0$. The interpretation of the acceleration $a = \nabla_u u$ is clear; by definition it vanishes for congruences of geodesics, for which

$$k_{\mu\nu} = \nabla_\mu u_\nu. \tag{6.20}$$

Furthermore, eq. (6.4) implies that h^μ_ν is the projection onto the orthogonal complement of u, since we have $h^\mu_\nu u^\nu = 0$, $h^\mu_\nu X^\nu = X^\mu$ whenever $g(u,X) = 0$, and finally, $h^\mu_\nu h^\nu_\rho = h^\mu_\rho$.

If u comes from a spacelike hypersurface Σ as explained above, the tensor $h_{\mu\nu}$ is a four-dimensional "covariant" version of the three-dimensional induced metric in Σ_t, in that

$$h_{\mu\nu} = g(h^\rho_\mu \partial_\rho, h^\sigma_\nu \partial_\sigma) = h^\rho_\mu h^\sigma_\nu g_{\rho\sigma}. \tag{6.21}$$

Proposition 6.3 *Let u be a timelike vector field in $U \subset M$ as above. Then there exists a spacelike surface $\Sigma \subset U$ to which the vectors u are normal iff $\omega_{\mu\nu} = 0$ (i.e. $k_{\mu\nu} = k_{\nu\mu}$).*

Proof. This follows from the Frobenius theorem, in the following form: the vectors orthogonal to u form an integrable distribution (that is, they span the tangent space to a (hyper)surface Σ orthogonal to u) iff they close under the Lie bracket. Here, this means that the condition

$$g(u,X) = g(u,Y) = 0, \tag{6.22}$$

implies $g(u,[X,Y]) = 0$. By (3.47) this is the same as $g(u, \nabla_X Y) = g(u, \nabla_Y X)$, or as

$$g(\nabla_X u, Y) = g(\nabla_Y u, X), \tag{6.23}$$

assuming (6.22), since $0 = X(g(u,Y)) = g(\nabla_X u, Y) + g(u, \nabla_X Y)$, and similarly with X and Y swapped. But (6.23), given (6.22), is equivalent to $k_{\mu\nu} = k_{\mu\nu}$ and hence to $\omega_{\mu\nu} = 0$. □

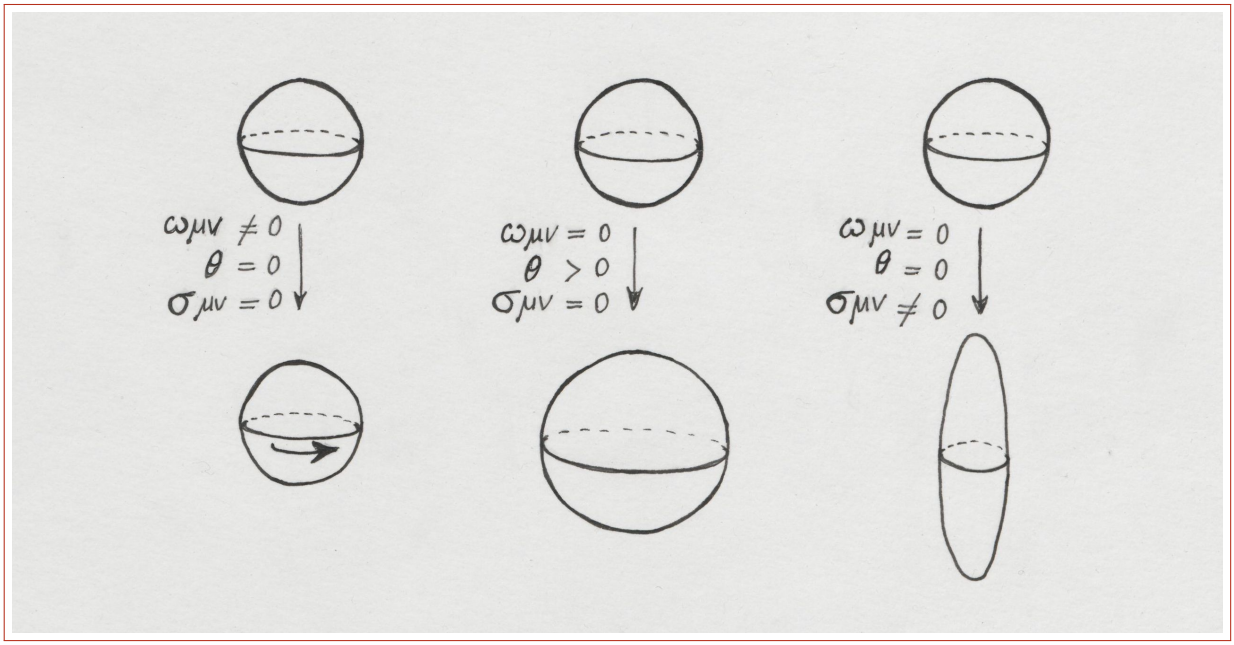
A good way to look at θ follows from the computation (7.25) in §7.1 below, in which ∂V should be replaced by Σ_t and $\vec{N}$ by u. Using (local) coordinates (x^1,x^2,x^3) on Σ_t, such that $ux^i = 0$ for each $i = 1,2,3$, the geometric volume form on Σ_t is given by

$$\sigma(x) := V(x)dx^1 \wedge dx^2 \wedge dx^3. \tag{6.24}$$

Then (7.25), which is $\mathscr{L}_u\sigma = \theta\sigma$, comes down to $V^{-1}\partial_t V = \theta$ or

$$\theta = \frac{1}{V}\frac{\partial V}{\partial t} = \partial_t \ln(V). \tag{6.25}$$

The terminology used in (6.10) - (6.15) is hybrid: the geometric notion of extrinsic curvature only really makes sense if the congruence arises from a hypersurface Σ in the said way, whereas the other terms rather come from fluid mechanics. The vorticity tensor describes the rotation of the fluid, the shear (which is traceless) describes the directed volume-preserving expansion (or, if negative, the contraction) of the fluid, and finally θ gives the rate of total volume increase (or, if negative, the decrease) under the flow. This is shown in the following picture: [285]



Left to right: effects of rigid rotation, uniform spherical expansion, and volume-preserving shear.

We finally derive the fundamental ***Raychaudhuri equation*** for θ.[286] Using (4.13), we compute

$$\begin{aligned} u^\sigma\nabla_\sigma(\nabla_\mu u_\nu) &= u^\sigma(\nabla_\mu\nabla_\sigma + [\nabla_\sigma,\nabla_\mu])u_\nu \\ &= \nabla_\mu(u^\sigma\nabla_\sigma u_\nu) - (\nabla_\mu u^\sigma)\nabla_\sigma u_\nu + R_{\nu\rho\sigma\mu}u^\sigma u^\rho. \end{aligned} \tag{6.26}$$

For geodesics the first term vanishes. Eqs. (6.15), (6.20), and (6.16) then yield, along u,

$$\nabla_u\theta \equiv \dot\theta = -\tfrac{1}{3}\theta^2 - \sigma_{\mu\nu}\sigma^{\mu\nu} + \omega_{\mu\nu}\omega^{\mu\nu} - R_{\mu\nu}u^\mu u^\nu. \tag{6.27}$$

This equation acts as a key lemma to Hawking's singularity theorem, to which we now turn.

[285] Redrawn from Malament (2012), p. 174, by Edith de Jong. Caption also due to Malament.

[286] Amal Kumar Raychaudhuri (1923–2005) was an Indian physicist. The July 2007 issue of *Pramana–Journal of Physics* (Volume 69, no. 1) is devoted to him and his equation. See also Earman (1999) for its historical context.

6.2 Hawking's singularity theorem

Hawking's singularity theorem from 1965–1966 remains exemplary because of the clarity of its hypotheses, its spectacular conclusion, and the elegance of its proof. Here it is:

Theorem 6.4 *Let (M,g) be a space-time. Assume that:*[287]

1. *(M,g) is globally hyperbolic;*
2. *One has $R_{\mu\nu}\dot\gamma^\mu\dot\gamma^\nu \geq 0$ along all timelike geodesics γ in M (***timelike curvature condition***);*
3. *The mean extrinsic curvature H of some spacelike Cauchy surface Σ (defined with respect to* future directed *timelike normals) satisfies $H(x) < H_0$ for some $H_0 < 0$ and all $x \in \Sigma$.*

Then (M,g) contains incomplete timelike geodesics: specifically, no past directed timelike geodesic γ emanating from Σ can have arc length (i.e. proper time) $L(\gamma) > 3/|H_0|$.

Before discussing the detailed meaning of the assumptions made here (including the notation H), let us state their generic nature, which is common to all singularity theorems in GR so far:[288]

1. is a global *causality* assumption;
2. is a *dynamical* assumption on the curvature motivated by a corresponding assumption on the energy-momentum tensor $T_{\mu\nu}$ in the Einstein equations (see below);
3. is a *static* assumption on the curvature, i.e. a boundary condition imposed at some fixed time, typically empirically motivated (in this case, by the expansion of the universe).

We have amply discussed global hyperbolicity: is it the strongest generic causality assumption. Logically speaking, strong assumptions weaken theorems. But since Hawking's theorem merely states that (M,g) is timelike incomplete without giving any indication why that is, global hyperbolicity at least strengthens the inference from geodesic incompleteness to (M,g) having some "interesting" singularity, see the discussion preceding Lemma 5.29 in §5.7.

As the proof will show, assumption 2 means that observers moving on timelike geodesics see these converge (by tidal forces), so that gravity is attractive. The Einstein equations

$$R_{\mu\nu} = 8\pi(T_{\mu\nu} - \tfrac{1}{2}g_{\mu\nu}T), \tag{6.28}$$

where $T_{\mu\nu}$ is the energy-momentum tensor (discussed in more detail in §7.3), relate assumption 2 to the matter content of the universe, notably to the so-called ***Strong Energy Condition (SEC)***

$$T_{\mu\nu}\dot\gamma^\mu\dot\gamma^\nu \geq -\tfrac{1}{2}T, \tag{6.29}$$

where $T = g^{\mu\nu}T_{\mu\nu}$. Writing $\dot\gamma = u$, the energy relative to our observer is $E = T_{\mu\nu}u^\mu u^\nu$ and the average (spatial) stresses are $S = g^{\mu\nu}h^\rho_\mu h^\sigma_\nu T_{\rho\sigma}$, where h is defined by (6.11). The inequality (6.29) then simply becomes $E \geq -S$, which is satisfied by most classical forms of matter.[289]

[287] We state the physically relevant case: there is a mathematically equivalent version assuming $H > 0$, in which case the *future* directed timelike geodesics will be incomplete (this describes a big crunch instead of a big bang).

[288] See Senovilla (1998) for a detailed introduction and overview, included the structure laid out here.

[289] See §7.3 for a brief discussion of energy conditions in GR, as well as Curiel (2014a) and Martín-Moruno & Visser (2017) for extensive information. As Curiel notes, '[SEC] says that ordinary mass-energy density cannot be negatively dominated by the sum of the individual pressures (momentum fluxes) at any point, as determined by an observer traversing a timelike curve. I know of no compelling elucidation of the physical content of that relation.'

The ***mean extrinsic curvature*** of $\Sigma \subset M$ was already introduced in §4.3, see (4.56), though in one dimension less and and in the special case where $M = \mathbb{R}^3$ with Euclidean metric. The general case was taken up in §4.7. We recall that first, the ***extrinsic curvature*** of $\Sigma \subset M$ is a tensor field $\tilde{k} \in \mathfrak{X}^{(2,0)}(\Sigma)$ initially defined merely on Σ by

$$\tilde{k}(X,Y) := -g(\nabla_X u, Y), \tag{6.30}$$

where $X, Y \in \mathfrak{X}(\Sigma)$ and u is the fd normal vector field on Σ discussed in the previous section. This definition is predicated on the fact that $\nabla_X u$ is tangent to Σ, which is an easy consequence of the property $g(u,u) = -1$. Similarly, it is easy to show that $\tilde{k}$ is symmetric, namely:

$$\tilde{k}(X,Y) = -g(\nabla_X u, Y) = g(u, \nabla_X Y) = g(u, \nabla_Y X) = \tilde{k}(Y,X). \tag{6.31}$$

The tensor $k_{\mu\nu}$ is just a covariant version of $\tilde{k}$, in that $k(u,u) = k(u,X) = \tilde{k}(u,u) = \tilde{k}(u,X) = 0$ whenever $g(u,X) = 0$, and $k(X,Y) = \tilde{k}(X,Y)$ if $g(u,X) = 0$ and $g(u,Y) = 0$, where $k(A,B) = k_{\mu\nu} A^\mu B^\mu$. From $\tilde{k}$, we define the ***mean (extrinsic) curvature*** $H : \Sigma \to \mathbb{R}$ of Σ as

$$H(x) := \mathrm{tr}(\tilde{k}_x) = \sum_{i=1}^{2} \tilde{k}_x(e_i(x), e_i(x)), \tag{6.32}$$

where $(e_i(x))$ is any orthogonal basis of $T_x\Sigma$, $x \in \Sigma$. Since k in (6.12) is spatial (because of the projections h in its definition), because of the (conventional) minus sign in (6.30), on Σ we have

$$\theta = -H. \tag{6.33}$$

Let us recall some Riemannian examples from §4.3, see (4.66), (4.70), and (4.74):

- For any plane in $\mathbb{R}^3$ (with flat metric) we have $H = 0$, and hence $\theta = 0$.
- For the cylinder of radius ρ, i.e., $C^2_\rho \subset \mathbb{R}^3$ (with flat metric), we have $\theta = 1/\rho$.
- For the sphere of radius ρ, i.e., $S^2_\rho \subset \mathbb{R}^3$ (with flat metric), we have $\theta = 2/\rho$.

Here the normal vectors used in the definition (6.30) are *outward*, and we see from these examples that negative H, and hence *positive* θ, gives diverging geodesics normally emanating from Σ. By the same token, *negative* θ gives *converging* normal geodesics, which in our universe happens in the *past* direction, so we have $H > 0$ on Σ in the *past* direction and hence $H < 0$ in the *future* direction, as assumed in the theorem (where $\dot{\gamma} = u$ is taken to be future directed). *The* Lorentzian example is the FLRW universe (cf. §8.3), where (8.88) gives $\theta(t) = 3\dot{a}(t)/a(t)$.

Proof. Given assumption 3, we can work in the setting explained after (6.4) in the previous section. We write $\dot{\gamma} = u$ and return to the Raychaudhuri equation (6.27). Since $\sigma_{\mu\nu}$ is symmetric and spatial, we have $\sigma_{\mu\nu}\sigma^{\mu\nu} = \mathrm{Tr}(\sigma^2) \geq 0$, where σ is the matrix with components $\sigma^\mu_\nu = h^{\mu\rho}\sigma_{\rho\nu}$. Furthermore, $\omega = 0$ by Proposition 6.3, whilst assumption 2 in Theorem 6.4 gives

$$R_{\mu\nu} u^\mu u^\nu \geq 0. \tag{6.34}$$

Therefore, the Raychaudhuri equation (6.27) gives $\dot{\theta} + \frac{1}{3}\theta^2 \leq 0$, i.e. $\dot{\theta}^{-1} \geq \frac{1}{3}$. Assumption 3 gives $\theta > \theta_0 > 0$ on Σ, where $\theta_0 = -H_0$, whence $0 < \theta^{-1} < \theta_0^{-1}$. It then follows that $\theta^{-1} \to 0$, or $\theta \to \infty$, at some time $t_s \in [3\theta_0^{-1}, 0)$, i.e., backward in time, *provided that the geodesic in question can indeed be extended to* t_s. By (6.25), this corresponds to $V \to 0$.

We now transform the divergence of θ into a conclusion about conjugate points. Within the congruence we take some fixed geodesic $\gamma\colon[-a,0]\to M$ with $\dot\gamma(t)=u(t)$ for all t and $\gamma(0)=x\in\Sigma$. As in §5.1, define a smooth family of neighbouring geodesics (γ_s) *all lying in the same congruence*, that is, $\gamma_s(0)\in\Sigma$ and $\dot\gamma_s(t)\perp u(t)=0$ for all s and t, and define the associated Jacobi field J by (5.10). Since $\dot\gamma=u$, we may then rewrite (5.5) as

$$\nabla_u J=\nabla_J u. \tag{6.35}$$

Since $\dim(\Sigma)=3$, the space $\mathsf{J}_\gamma^{(\Sigma)}\subset\mathsf{J}_\gamma$ of Jacobi fields arising in this way is 3-dimensional. Hence it is convenient to introduce a moving frame $(e_1(t),e_2(t),e_3(t))$ along $\gamma(t)$, that is, an orthonormal basis of $T^\perp_{\gamma(t)}M$ (see Proposition 5.2) for each t; such a frame can be constructed by solving $\nabla_{\dot\gamma}e_i=\nabla_u e_i=0$ with orthonormal initial conditions at $t=0$; this equation guarantees that the frame remains orthonormal as well as orthogonal to $\dot\gamma$.[290] We may then expand

$$J(t)=J_i(t)e_i(t)\equiv\sum_{i=1}^3 J_i(t)e_i(t); \tag{6.36}$$

$$J_i(t)=g_{\gamma(t)}(J(t),e_i(t)). \tag{6.37}$$

Furthermore, the covariant derivative $\nabla_{\dot\gamma}J=\nabla_u J$ now becomes a time-derivative, since

$$(\nabla_u J)_i(t)=g_{\gamma(t)}(\nabla_u J(t),e_i(t))=u(g_{\gamma(t)}(J(t),e_i(t)))-g_{\gamma(t)}(J(t),\nabla_u e_i(t))=\dot J_i(t), \tag{6.38}$$

where $\dot J_i(t)=dJ_i(t)/dt$. Using (6.35) and $\nabla_{\dot\gamma}e_i=0$, we obtain

$$\begin{aligned}\dot J_i(t)&=g_{\gamma(t)}(\nabla_u J(t),e_i(t))=g_{\gamma(t)}(\nabla_J u(t),e_i(t))=J_j(t)g_{\gamma(t)}(\nabla_j u(t),e_i(t))\\&=k_{ij}(t)J_j(t),\end{aligned} \tag{6.39}$$

where, keeping in mind that $k_{ij}=-\tilde k_{ij}$, cf. the minus sign in (6.30),

$$k_{ij}(t)=g_{\gamma(t)}(\nabla_j u(t),e_i(t))=k_{\mu\nu}(\gamma(t))e_i^\mu(t)e_j^\nu(t), \tag{6.40}$$

are the components of (6.30) in the frame $(e_i(t))$, with $k_{ij}=k_{ji}$. Eq. (5.7) then reads

$$\frac{d^2J_i(t)}{dt^2}=\alpha_{ij}(t)J_j(t), \tag{6.41}$$

$$\alpha_{ij}:=g(e_i,\Omega(u,e_j)u)=R(e_i,u,e_j,u). \tag{6.42}$$

Conversely, a simple dimension count shows that the Jacobi fields $J\in\mathsf{J}_\gamma^{(\Sigma)}$ along γ that arise from the congruence emanating from Σ in the said way are those that satisfy the initial condition

$$\dot J_i(a)=k_{ij}(a)J_j(a), \tag{6.43}$$

predicated on (6.36) - (6.37), so that (6.43) implicitly also assumes the initial condition

$$J(a)\in T_{\gamma(a)}\Sigma\ \Leftrightarrow\ g_{\gamma(a)}(J,u)=0. \tag{6.44}$$

Conversely, it follows from the Jacobi equation (6.41) with (6.42) that both initial conditions (6.39) and (6.44) are propgated along γ, i.e., hold for all t where $\gamma(t)$ is defined.

For our proof we now need the following variation on Definition 5.10, backward in time.

[290] This simple construction works because γ is a geodesic. Along more general curves one needs the so-called ***Fermi derivative*** $\nabla^F_{\dot\gamma}e_i$ instead of the covariant derivative $\nabla_{\dot\gamma}e_i$. See e.g. Hawking & Ellis, §4.1.

Definition 6.5 *A point $x=\gamma(c)$, where $c\in[-a,0)$, is* **focal** *relative to $y=\gamma(0)$, seen as a member of the congruence of normal geodesics to Σ, if there is a nonzero Jacobi field $J\in \mathsf{J}^{(\Sigma)}_\gamma$, i.e. satisfying the Jacobi equation along γ with initial condition (6.43), for which $J(c)=0$.*

To elucidate this, for any $t\in[-a,0]$ we define two "double evaluation maps" A_t, B_t by

$$A_t, B_t : \mathsf{J}_\gamma \to T^\perp_{\gamma(0)}M \oplus T^\perp_{\gamma(t)}M; \tag{6.45}$$

$$A_t(J)=(J(0),J(t)); \qquad B_t(J)=(\nabla_t J(0)-k_{\gamma(0)}J(0),J(t)). \tag{6.46}$$

Both maps are linear, and we see that $\gamma(c)$ is conjugate relative to $\gamma(0)$ iff A_c is singular, whereas it is focal iff B_c is singular. Despite this difference, Theorem 5.12 applies, *mutatis mutandis*:

Theorem 6.6 *A timelike geodesic $\gamma:[-a,0]\to M$ in a congruence as above locally maximizes the length of (pd) curves from $\gamma(0)$ to $\gamma(-a)$ iff there is* no *focal point $\gamma(c)$ on γ, $c\in[-a,0]$.*

The proof is the same as for Theorem 5.12, since Synge's formula (5.21) still holds. This is remarkable in itself, since in its derivation one now picks up boundary terms at a, because this time $J(0)\neq 0$. There are two: first, after (5.18) one needs to add $-g_{\gamma(0)}(\nabla_s\gamma',\dot\gamma)$, which equals

$$-g_{\gamma(0)}(\nabla_s\gamma',\dot\gamma)=-\frac{d}{ds}g_{\gamma(0)}(\gamma',\dot\gamma)+g_{\gamma(0)}(\gamma',\nabla_s\dot\gamma)=g_{\gamma(0)}(\gamma',\nabla_t\gamma')=g_{\gamma(0)}(J,\nabla_t J), \tag{6.47}$$

since in our case $g_{\gamma(0)}(\gamma',\dot\gamma)=g_{\gamma(0)}(J,u)=0$, and we also used (5.5). Second, after (5.20) one picks up $-g_\gamma(\gamma'_\perp,\nabla_t\gamma'_\perp)=-g_{\gamma(0)}(J,\nabla_t J)$, which fortunately cancels the term in (6.47).

We now relate the existence of focal points to the expansion θ of the congruence. It follows from Proposition 5.2, which makes $J(t)$ depend linearly on the initial conditions $J(0)$ and $\dot J(0)$, and eq. (6.43), according to which $\dot J(0)$ depends linearly on $J(0)$, that if $J\in\mathsf{J}^{(\Sigma)}_\gamma$, then

$$J_i(t)=A_{ij}(t)J_j(0)\ \ (t\in[-a,0]), \tag{6.48}$$

for some 3×3 matrix $A(t)$. From (6.39) and(6.48) we obtain

$$\dot J^i(t)=\dot A_{ij}(t)J^j(0)=k_{ij}(t)J^j(t)=k_{ij}(t)A_{jk}(t)J^k(0), \tag{6.49}$$

so that $\dot A_{ik}=k_{ij}A_{jk}$, or $k=\dot A A^{-1}$, and hence, since $\theta=\mathrm{tr}(k)=-\mathrm{tr}(\tilde k)$, we finally obtain

$$\theta=\mathrm{tr}(\dot A A^{-1}). \tag{6.50}$$

Now A is finite along γ, and so is $\dot A$. Hence in the scenario just considered, θ can only blow up at $t=c$ iff $A(c)^{-1}$ does, i.e., iff $A(c)$, which equals the identity at $t=0$, has an eigenvalue zero. But as explained after (6.46), this implies that there exists some $J\in\mathsf{J}^{(\Sigma)}_\gamma$ for which $J(c)=0$, which means that $\gamma(c)$ is a focal point with respect to $\gamma(0)$. So if $\theta(\gamma(0))>0$, then $\gamma(c)$ is a focal point with respect to $\gamma(0)$ iff $\lim_{t\to c}\theta(t)=\infty$. Therefore, the argument after (6.34) gives:

Proposition 6.7 *Let γ be an element of the congruence of timelike geodesics orthogonal to a spacelike hypersurface $\Sigma\subset M$. If the positive curvature condition (6.34) holds along γ, and if $\theta(\gamma(0))>0$ somewhere along γ, then γ has an earlier focal point $\gamma(c)$ relative to $\gamma(0)$, provided that the geodesic in question can indeed be extended (backward) from $t=0$ all the way to $t=c$.*

Similarly, we need a corollary to Theorem 5.30 describing the case where x is connected to Σ (rather than to a given single point y) by a length-maximizing timelike geodesic.

Corollary 6.8 *Let $\Sigma \subset M$ be a spacelike Cauchy hypersurface. For any $x \in I^-(\Sigma)$ there is a (not necessarily unique) future-directed timelike geodesic from x to Σ that maximizes length among all timelike geodesic from x to Σ. This geodesic necessarily crosses Σ orthogonally.*

Proof. It is easy to see that $J^+(x) \cap \Sigma$ is compact. Indeed, even $J^+(x) \cap D^-(\Sigma)$ is compact.[291] Furthermore, if (M,g) is globally hyperbolic, then the Lorentzian distance function d_L defined by (5.118) is continuous in both arguments.[292] For any $y \in J^+(x) \cap \Sigma$, Theorem 5.30 gives a length-maximizing causal geodesic from x to y whose length equals $d_L(x,y)$, so keeing x fixed we have a continuous function of y that assumes a maximum on the compact set $J^+(x) \cap \Sigma$, say at y_0. Since $x \in I^-(\Sigma)$, the maximizing geodesic γ from x to y_0 must be timelike by Proposition 5.13. Finally, the boundary term $|_b^a g(\gamma', \dot\gamma)$ in (5.15), which has to vanish, shows that γ crosses Σ orthogonally, since the variation γ' vanishes at $a = 0$ and is tangent to Σ at b. □

At last, we are now in a position to prove Hawking's Theorem 6.4. It is sufficient to prove it for timelike geodesics *normally* emanating from Σ, since by Corollary 6.8 timelike geodesics not starting normally from Σ are shorter than those which do. The proof is by contradiction.

1. Take some $x \in I^-(\Sigma)$ and a length-maximizing timelike geodesic γ from x to $y_0 \in \Sigma$, as in the proof of Corollary 6.8. Suppose $L(\gamma) > 3/|H_0|$.

2. Since γ crosses Σ orthogonally, it is a member of the congruence described in the previous section, and so by Proposition 6.7, γ will have focal points (backward in time).

3. By Theorem 6.6, the length-maximixing γ cannot have any focal points.

4. Hence γ cannot exist: it must have stopped before $t = -3/|H_0|$. Hence it is incomplete and the contradiction is resolved because γ never reaches its would-be focal point.

5. Any $y \in \Sigma$ can be reached in this way, and as already noted, the conclusion that normal timelike geodesics are incomplete implies the same conclusion for any timelike geodesic starting at Σ, with the same bound (since they are shorter than the normal ones). □

Here assumption 1 in the theorem (i.e. global hyperbolicity) is used (via Corollary 6.8) in step 1, whereas the two curvature assumptions are exploited in step 2. Note that the time to reach the singularity *increases* as the mean extrinsic curvature Σ *decreases*, in accordance with intuition: less curvature means less focusing (and a higher age of the universe). Of course, timelike geodesic incompleteness can still be proved if the uniform bound on the extrinsic curvature is replaced by local bounds, but the stated version of Hawking's theorem is meant to describe the big bang, where *every* indextendible past-directed timelike geodesic ends.[293]

[291] See O'Neill (1983), Lemma 14.40 for the even more general statement that for any achronal set A and $x \in \mathrm{int}(D(A)) \backslash I^+(A)$, the set $J^+(x) \cap D^-(A)$ is compact. Correct intuition is obtained by taking a future inextendible fd timelike curve c from x crossing Σ. Then $J^-(c(t))$ eventually covers all of M as $t \to \infty$, so that there is t_0 such that $J^+(x) \cap \Sigma \subset J^-(c(t_0))$. Hence $J^+(x) \cap \Sigma \subset J^-(c(t_0)) \cap J^+(x)$, which is compact by global hyperbolicity. Since Σ and $J^+(x)$ are closed, this implies that $J^+(x) \cap \Sigma$ is a closed subset of a compact set, so that it is compact.

[292] See, with increasing eye for detail, Hawking (1966/2014), pp. 482–483, O'Neill (1983), Lemma 14.21, and Minguzzi (2019), Theorems 3.48 and 4.124. The key point is that d_L is always lower semicontinuous, upon which global hyperbolicity also makes it upper semicontinuous, using upper semicontinuity of $L(\cdot)$ and Lemma 5.26.

[293] Theorem 5 in Hawking (1966/2014) is a singularity theorem under weaker conditions, keeping the two curvature assumptions but replacing global hyperbolicity by the mere existence of a *compact* spacelike hypersurface. The version above is Hawking' Theorem 3. See also O'Neill (1983), Theorems 14.55A and 14.55B.

6.3 Null congruences and trapped surfaces

This section prepares for Penrose's singularity theorem from 1965. Whereas Hawking's theorem is based on timelike geodesics and more generally on 'timelike reasoning', Penrose's theorem (indeed his entire approach to GR, cf. §1.9), is based on 'lightlike reasoning', based on null geometry. This requires a preamble, which we started in §4.6 and now continue.[294]

Proposition 6.9 *If $\Sigma \subset M$ is a null hypersurface with normal null vector field*

$$x \mapsto L_x \in T_x\Sigma \cap (T_x\Sigma)^\perp \ \ (x \in \Sigma); \qquad\qquad g(L,L) = 0, \tag{6.51}$$

then the flow lines of L are lightlike pregeodesics and hence Σ is ruled by lightlike geodesics.[295]

To prove this, for any $X \in T\Sigma$, so that $g(X,L) = 0$, we compute (omitting the subscript $x \in \Sigma$):

$$\begin{aligned} &0 = Lg(X,L) = (\nabla_L g)(X,L) + g(\nabla_L X, L) + g(X, \nabla_L L) = g(\nabla_L X, L) + g(X, \nabla_L L); \\ &\Rightarrow \ g(\nabla_L L, X) = -g(L, \nabla_L X) = -g(L, \nabla_X L) - g(L, [L,X]) = 0, \end{aligned} \tag{6.52}$$

where we used torsionlessness of the Levi-Civita connection ∇, as well as the computations

$$g(L, \nabla_X L) = \tfrac{1}{2} X g(L,L) = 0; \tag{6.53}$$

$$g(L, [L,X]) = g(L, \mathscr{L}_L X) = 0, \tag{6.54}$$

as follows from (3.50) and (2.35), respectively; if $L \in T\Sigma$ and $X \in T\Sigma$, then also $\mathscr{L}_L X \in T\Sigma$ and hence this vector is orthogonal to L. Eq. (6.52) implies that $\nabla_L L$ is orthogonal to every vector X tangent to Σ, and hence must be proportional to its normal L. By Proposition 3.8, the flow of L may therefore be (re)parametrized so as to be geodesic. See also Theorem 5.5.2. □

As a special case, consider a hypersurface $\Sigma = \{u = c\}$ (locally) defined by a smooth function u, where $c \in \mathbb{R}$ (or rather a family thereof). Then $N \equiv L = \nabla u$, so if Σ is null, then u satisfies

$$g(\nabla u, \nabla u) = 0. \tag{6.55}$$

This is the basic ***eikonal equation*** of hyperbolic PDEs, and u is called an ***optical function***. For example, the coordinate functions $u = t - r$ and $v = t + r$ on Minkowski space-time are optical.

Lemma 6.10 *If u is an optical function, then the flow of $L = \nabla u$ consists of lightlike geodesics.*

From $\nabla_\nu \partial_\mu u = \nabla_\mu \partial_\nu u$ for any function u (since ∇ is torsion-free), for any vector field X,

$$X^\nu L^\mu \nabla_\nu \partial_\mu u = X^\nu L^\mu \nabla_\mu \partial_\nu u, \tag{6.56}$$

i.e. $g(L, \nabla_X L) = g(X, \nabla_L L)$, where $L = \nabla u$. By (6.53), which is true for any X (not necessarily in $T\Sigma$) as long as $g(L,L) = 0$, this gives

$$g(X, \nabla_L L) = 0 \tag{6.57}$$

for any X, whereas the previous lemma merely showed this for $X \in T\Sigma$. Hence $\nabla_L L = 0$. Thus the flow of L consists of lightlike geodesics (without the need for reparametrization). □

[294] Kupeli (1987), Aretakis (2013), and Galloway (2014, 2017) are useful introductions to null geometry.

[295] I.e. a unique lightlike geodesic passes through every point of Σ. Conversely, a hypersurface is null if it is ruled by lightlike geodesics *and locally achronal*, see Kupeli (1987), Theorem 1 and Minguzzi (2019), Theorem 6.7.

A ***section*** of a null hypersurface Σ is a two-dimensional surface $S \subset \Sigma$ such that T_xS is spacelike (i.e. Riemannian) for each $x \in S$; hence $T_x\Sigma = T_xS \oplus \mathbb{R} \cdot L$. In 2+1-dimensional Minkowski space-time, sections of forward or backward lightcones are circles and, less easily visualized, in $d = 3+1$ they are two-spheres. Conversely, start from an oriented closed surface $S \subset M$ (i.e. S is $2d$, compact, and without boundary–in this context one may think of the two-sphere S^2). At each $x \in S$, the orthogonal complement $(T_xS)^\perp$ has signature $(-+)$ and hence is spanned by two future-directed lightlike vectors L_x and $\underline{L}_x$. These lightlike vectors may be normalized by

$$g_x(L_x, \underline{L}_x) = -2, \tag{6.58}$$

and together with any basis (e_1, e_2) of T_xS they form a basis of T_xM. For example, the pair

$$L := 2\partial_v = \partial_t + \partial_r; \tag{6.59}$$

$$\underline{L} := 2\partial_u = \partial_t - \partial_r. \tag{6.60}$$

does the job in Minkowski space-time $(\mathbb{M}, \eta)$. In that case, the family (L_x) is directed outward and diverges off to infinity, whereas the other one, viz. $(\underline{L}_x)$, is directed inward and converges to an apex like a Chinese hat. But in general space-times this may not be the case; e.g. inside a black hole both families bend inwards and one has a *trapped surface* (cf. Definition 6.13).

At any $x \in S$, consider the fd lightlike geodesic $\gamma_L^{(x)}$ with $\gamma^L_{(x)}(0) = x$ and $\dot\gamma_L^{(x)}(0) = L_x$. These geodesics collectively form a ***null congruence*** emanating from S, that is, a hypersurface

$$C := \bigcup_{x\in S}\bigcup_{t\geq 0} \gamma_L^{(x)}(t) \equiv \bigcup_{t\geq 0} S_t, \tag{6.61}$$

where

$$S_t := \bigcup_{x\in S} \gamma_L^{(x)}(t) \tag{6.62}$$

is the image of $S = S_0$ at time t under the geodesic flow in question, as long as it is defined. Note that this is not really a hypersurface as we defined it, because it has a boundary

$$\partial C = S. \tag{6.63}$$

Minkowski space-time shows that C may develop conical singularities at finite t and hence may be a (smooth) surface only up to some t_f. In that case, except for the boundary (6.63):

Proposition 6.11 *The set C defined by (6.61) is a null hypersurface as long as it is smooth.*

First, the lightlike vector field L may be extended from S to C in the obvious way, namely by

$$L_{\gamma_L^{(x)}(t)} = \dot\gamma_L^{(x)}(t), \tag{6.64}$$

so that, since each $\gamma_L^{(x)}$ is a geodesic, $\nabla_L L = 0$. If we push forward any $X_x \in T_xS$ to $T_{\gamma_L^{(x)}(t)}C$ by the flow of L, then the ensuing vector field X along $\gamma_L^{(x)}$ satisfies $\mathscr{L}_L X = [L, X] = 0$, so that

$$\begin{aligned}
\frac{d}{dt} g_{\gamma_L^{(x)}(t)}(L_{\gamma_L^{(x)}(t)}, X_{\gamma_L^{(x)}(t)}) &= L g_{\gamma_L^{(x)}(t)}(L_{\gamma_L^{(x)}(t)}, X_{\gamma_L^{(x)}(t)}) \\
&= g_{\gamma_L^{(x)}(t)}((\nabla_L L)_{\gamma_L^{(x)}(t)}, X_{\gamma_L^{(x)}(t)}) + g_{\gamma_L^{(x)}(t)}(L_{\gamma_L^{(x)}(t)}, (\nabla_L X)_{\gamma_L^{(x)}(t)}) \\
&= 0 + g_{\gamma_L^{(x)}(t)}(L_{\gamma_L^{(x)}(t)}, (\nabla_X L)_{\gamma_L^{(x)}(t)}) \\
&= \tfrac{1}{2} X g_{\gamma_L^{(x)}(t)}(L_{\gamma_L^{(x)}(t)}, L_{\gamma_L^{(x)}(t)}) \\
&= 0.
\end{aligned} \tag{6.65}$$

Therefore, for fixed $x \in S$, the function

$$t \to g_{\gamma_L^{(x)}(t)}(L_{\gamma_L^{(x)}(t)}, X_{\gamma_L^{(x)}(t)}) \tag{6.66}$$

is constant and hence equal to its value at $t = 0$, i.e., at S, where it vanishes (since L is normal to S). Since $g(L,L) = 0$, this is true also for $X = L$, so that L is also orthogonal to C. This makes C (or rather $C \backslash S$) a null hypersurface by definition. □

Thus null hypersurfaces may either be constructed from optical functions or from spaceike 2-surfaces. The former is relevant to the Cauchy problem, whereas the latter applies to Penrose's singularity theorem, to which we now slowly turn. Compared to Hawking,'s the three-dimensional spacelike hypersurface Σ is replaced by a $2d$ closed spacelike surface S (with special properties to be defined), from which one proceeds as explained above:

- Construct a null hypersurface C with normal vector field L, where $S \subset C \subset M$;
- At each $x \in C$, construct a basis $(e_1, e_2, \underline{L}, L)$, normalized, also repeating (6.58), by

$$g(e_i, e_j) = \delta_{ij} \ \ (i,j = 1,2); \qquad g(e_i, L) = g(e_i, \underline{L}) = 0 \ \ (i = 1,2); \tag{6.67}$$

$$g(L,L) = g(\underline{L}, \underline{L}) = 0; \qquad g(L, \underline{L}) = -2. \tag{6.68}$$

This can be done by a slight refinement of the construction used for the spacelike case: Starting at $x \in S$, seen as the initial point of a lightlike geodesic $\gamma_L^{(x)}(\cdot) \equiv \gamma$ as above, and defining the basis at x, extend L and $\underline{L}$ as explained above, and extend (e_1, e_2) by solving

$$\nabla_L e_i = -g(\nabla_i L, \underline{L}) L. \tag{6.69}$$

The definition of Jacobi fields along γ obtained by varying γ within the congruence of all lightlike geodesics emanating from S, as in §6.3, is then entirely similar to the spacelike case. If

$$\mathscr{L}_L J = 0; \qquad g(J,L) = g(J, \underline{L}) = 0 \tag{6.70}$$

along γ, then $J(t) = \sum_{i=1}^{2} J_i(t) e_i(t)$ satisfies Jacobi's equation (6.41), this time with a matrix

$$\alpha_{ij} = g(e_i, \Omega(L, e_j) L) = R(e_i, L, e_j, L). \tag{6.71}$$

The computations leading to (6.38) and (6.39) may also be redone, *mutatis mutandis*:[296]

$$\begin{aligned}(\nabla_L J)_i(t) &= g_{\gamma(t)}(\nabla_L J(t), e_i(t)) = L(g_{\gamma(t)}(J(t), e_i(t))) - g_{\gamma(t)}(J(t), \nabla_L e_i(t)) \\ &= \dot{J}_i(t) + g_{\gamma(t)}(\nabla_i L(t), \underline{L}(t)) \cdot \sum_i J_i(t) g_{\gamma(t)}(e_i(t), L(t)) = \dot{J}_i(t); \\ \dot{J}_i(t) &= g_{\gamma(t)}(\nabla_L J(t), e_i(t)) = g_{\gamma(t)}(\nabla_J L(t), e_i(t)) = J_j(t) g_{\gamma(t)}(\nabla_j u(t), e_i(t)) \\ &= k_{ij}(t) J_j(t),\end{aligned} \tag{6.72}$$

where (6.40) is replaced by

$$k_{ij}(t) = g_{\gamma(t)}(\nabla_j L(t), e_i(t)). \tag{6.73}$$

Since C (or the S_t) is given, the distribution spanned by e_i and e_2 is integrable and Frobenius's theorem again gives $k_{ij} = k_{ji}$, where $i = 1,2$ instead of $i = 1,2,3$ as in §6.1, cf. Proposition 6.3.

For Penrose's singularity theorem we need the following variation on Definition 6.5, in which $\mathrm{J}_\gamma^{(S)}$ (replacing $\mathrm{J}_\gamma^{(\Sigma)}$ in Definition 6.5) denotes the space of Jacobi fields along γ satisfying (6.70), or, equivalently, (6.41) with (6.71), on the initial conditions (6.70) and (6.72) at $t = a$.

[296] One may introduce the form k on a null hypersurface Σ with normal L in a basis-independent way, namely as a bilinear form on $T_x\Sigma / K_x$, where K_x is the linear span of L_x ($x \in \Sigma$). The Lorentzian metric g, which is degenerate on $T_x\Sigma$, induces a nondegenerate (and Riemannian) metric h on $T_x\Sigma / K_x$, in terms of which $k([X],[Y]) = h([\nabla_X L],[Y])$.

Definition 6.12 *A point $y=\gamma(c)$, where $a<c\leq b$, is* **focal** *relative to $x=\gamma(a)$, seen as a member of the congruence of lightlike geodesics emanating from S and lying in C, if there is a nonzero Jacobi field $J\in \mathsf{J}^{(S)}_{\gamma}$ satisfying $J(c)=0$, as well as (6.70) and (6.72) at $t=a$.*

The smoothness of C breaks down at focal points. The simplest example is Minkowski space-time (best visualized in dimension $d=2+1$, where C is a spacelike circle): either follow the geodesics back in time, or use the ingoing lightlike $\underline{L}$ rather than L. All lightlike geodesics then have the same focal point. If we now define the ***null expansion*** θ on C by

$$\theta := \mathrm{tr}(k), \tag{6.74}$$

then the arguments leading from (6.48) to (6.50) may be repeated *verbatim*, and so it follows that $\gamma(c)$ is a focal point iff the scalar blows up at c. Before giving conditions for this to happen, we first investigate its geometrical meaning. Take coordinates (x^1,x^2) on S (for example, if $S=S^2$, the usual angles $(x^1=\varphi,x^2=\theta)$, so that the volume of S_t is given by

$$\mathrm{Area}(S_t) := \iint_{S_t} dx^1dx^2\,\sqrt{\det h_{\gamma(t)}(x^1,x^2)} \equiv \iint_{S_t} d\mu_t(x^1,x^2), \tag{6.75}$$

where $h_{\gamma(t)}$ is the metric on $S_t\subset M$ induced by the metric g on M, i.e. $h_{\gamma(t)}(X,Y)=g_{\gamma(t)}(X,Y)$ for $X,Y\in T_{\gamma(t)}S_t$. If d/dt is the directional derivative along L, the key formula, then, is

$$\frac{d\mathrm{Area}(S_t)}{dt} = \iint_{S_t} d\mu_t\,\theta(t), \tag{6.76}$$

so that θ measures the rate of change of the area of S_t when moving along the geodesic γ_L. In particular, this area shrinks when $\theta<0$ and decreases to zero at a focal point, where $\theta=-\infty$.

To prove (6.76), we interpret $k_{ij}(t)$ as defined in (6.73) as a (symmetric) bilinear map

$$k(t): T_{\gamma(t)}S_t\times T_{\gamma(t)}S_t\to\mathbb{R}; \qquad k(X,Y):=g(\nabla_Y L,X), \tag{6.77}$$

so that we may redefine θ by rewriting (6.74) with respect to an arbitrary coordinate basis as

$$\theta(t)=\sum_{I,J=1} h^{IJ}(t)k_{IJ}(t). \tag{6.78}$$

As usual, in this formula h^{IJ} is the inverse matrix to $h_{IJ}=h(\partial/\partial x^I,\partial/\partial x^J)$ for any (local) coordinates (x^1,x^2) on S_t. Using (3.73) with $X=L$, eq. (6.77), and the symmetry of k, yield

$$\mathscr{L}_L h_{IJ}(t)=\dot{h}_{IJ}(t)=2k_{IJ}(t), \tag{6.79}$$

upon which the elementary computation (7.12) and text below, applied to h, gives (6.76):

$$\frac{d}{dt}\sqrt{\det h_{\gamma(t)}}=\tfrac{1}{2}\sqrt{\det h_{\gamma(t)}}\,h^{IJ}(t)\dot{h}_{IJ}(t)=\sqrt{\det h_{\gamma(t)}},h^{IJ}(t)k_{IJ}(t)=\sqrt{\det h_{\gamma(t)}}\,\theta(t). \tag{6.80}$$

Exactly the same constructions apply to the null hypersurface $\underline{C}$ that is obtained from the given surface S by replacing L by $\underline{L}$ (and *vice versa*) throughout (6.61) - (6.76). For example, if in Minkowski space-time C is erected from the outgoing lightlike directions, then $\underline{C}$ is built from the ingoing ones, so that a focal point arises in the future as a Chinese hat. Thus we define

$$\underline{S}_{\underline{t}}=\bigcup_{x\in S}\gamma_x^{(\underline{L})}(\underline{t}); \qquad \underline{C}=\bigcup_{\underline{t}\geq 0}\underline{S}_{\underline{t}}; \tag{6.81}$$

$$\underline{k}(X,Y)=g(\nabla_Y\underline{L},X); \qquad \underline{\theta}=\mathrm{tr}(\underline{k}). \tag{6.82}$$

In Minkowski space-time the lightlike vectors (6.59) - (6.60) are easily shown to give

$$\theta(t,r,\theta,\varphi) = 2/r; \qquad \underline{\theta}(t,r,\theta,\varphi) = -2/r. \tag{6.83}$$

Assumption 3 on Σ in Hawking's Theorem 6.4 is now replaced by Penrose's assumption on S:

Definition 6.13 *A future* **trapped surface** *is a closed spacelike surface $S \subset M$ with*

$$\theta(x) < 0; \qquad \underline{\theta}(x) < 0 \tag{6.84}$$

for all $x \in S$, where θ and $\underline{\theta}$ are defined by (6.74) and (6.82), with L and $\underline{L}$ future directed.[297]

This "infinitesimally" states that *all* fd lightlike geodesics emanating orthogonally from S bend inwards, not just–as in Minkowski space-time– those along $\underline{L}$.[298] The picture on the next page illustrates this. The signs of θ and $\underline{\theta}$ depend on the time-orientation of the lightlike vectors L and $\underline{L}$; there is an analogous concept of *past* trapped surfaces. An appealing interpretation of Definition 6.13 follows from a slight generalization of (6.76). Let X be some convex combination of L and $\underline{L}$ and let $S_t^{(X)}$ be the image of S under the flow φ_t of X, $t > 0$, so that $S_t = S_t^{(L)}$. Then

$$\frac{d\text{Area}(S_t^{(X)})}{dt} = -\iint_{S_t} d\mu_t \, g(X, \theta(t)\underline{L}(t) + \underline{\theta}(t)L(t)), \tag{6.85}$$

so that, the area of a trapped surface S decreases in any future orthogonal direction.

Further changes from Hawking's setting to Penrose's are:

- The spacelike $(3d)$ hypersurface Σ is replaced by a closed spacelike surface S;
- For the normal u one may take either L or $\underline{L}$ (as both vectors are orthogonal to S);
- The expressions (6.10) - (6.16) now become

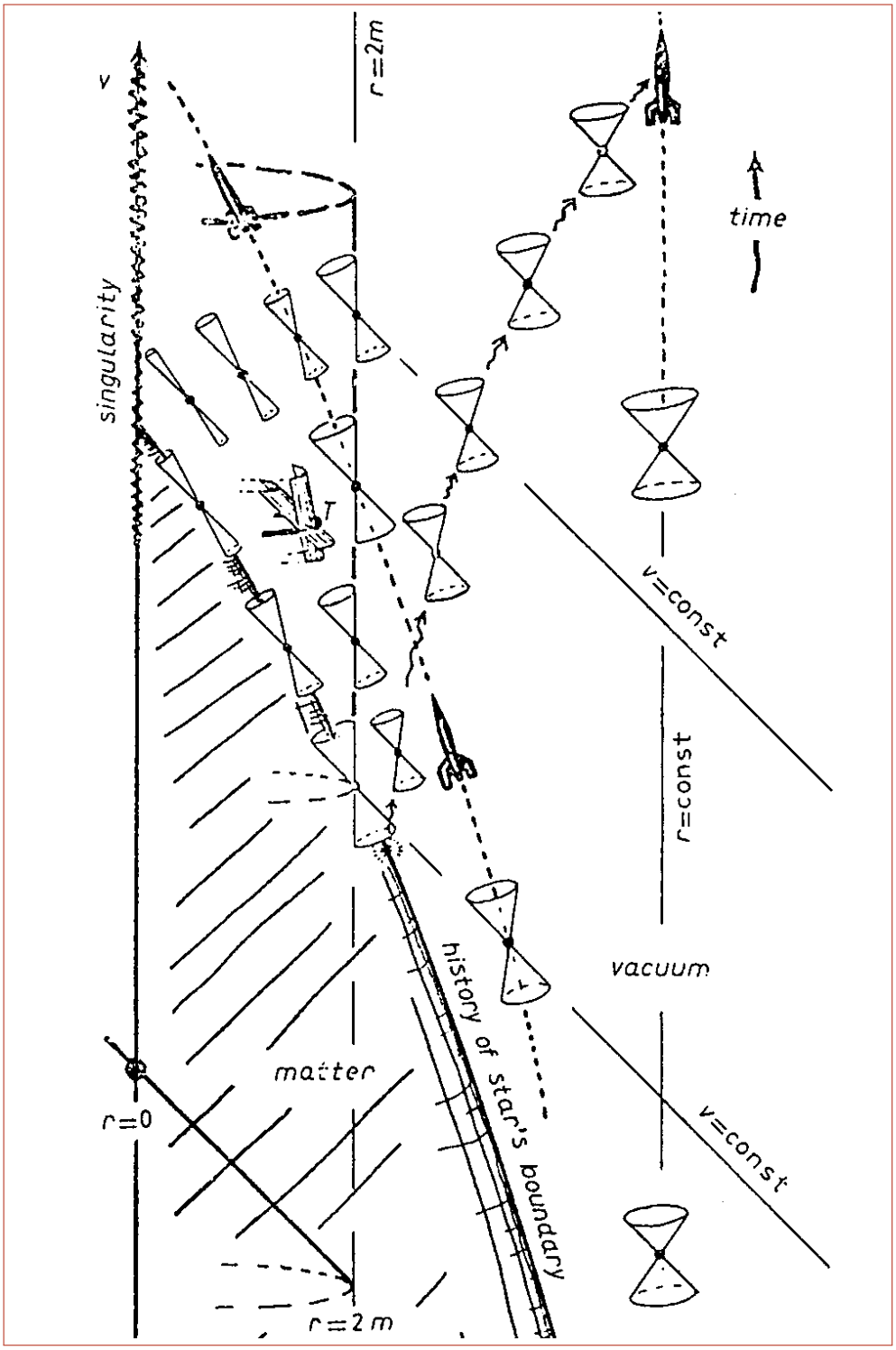
$$A^\mu := L^\nu \nabla_\nu L^\mu; \tag{6.86}$$
$$h^\mu_\nu := \delta^\mu_\nu + \tfrac{1}{2}(\underline{L}^\mu L_\nu + L^\mu \underline{L}_\nu); \tag{6.87}$$
$$k_{\mu\nu} := h^\rho_\mu h^\sigma_\nu \nabla_\rho L_\sigma; \tag{6.88}$$
$$\omega_{\mu\nu} := k_{[\mu\nu]}; \tag{6.89}$$
$$\sigma_{\mu\nu} := k_{(\mu\nu)} - \tfrac{1}{2}\theta h_{\mu\nu} \tag{6.90}$$
$$\theta := g^{\mu\nu}k_{\mu\nu} = h^{\mu\nu}k_{\mu\nu} \equiv \text{tr}(k); \tag{6.91}$$
$$\Rightarrow \; k_{\mu\nu} = \tfrac{1}{2}\theta h_{\mu\nu} + \sigma_{\mu\nu} + \omega_{\mu\nu}. \tag{6.92}$$

We have 1/2 in (6.90) and (6.92) as opposed to the 1/3 in (6.14) and (6.16) because σ is the traceless part of k (its trace already being taken care of by θ), and this time,

$$g^{\mu\nu}h_{\mu\nu} = \delta^\mu_\mu + \tfrac{1}{2}(g(\underline{L},L) + g(L,\underline{L})) = 4 - 1 - 1 = 2 = \dim(S). \tag{6.93}$$

[297] See Hawking & Ellis (1973), chapter 9, for an early analysis of trapped surfaces in causal theory *per se*. The study of trapped surface formation from the PDE point of view began with Schoen & Yau (1983), who gave initial values that *already contain* trapped surface; see also Alaee, Lesourd, & Yau (2019). Christodoulou (1991, 1999a, 2009) first proved the evolution of asymptotically flat initial data *into* trapped surfaces. See also follow-ups by Klainerman & Rodnianski (2012) and Klainerman, Luk, & Rodnianski (2014), and reviews by Dafermos (2012) and Bieri (2018). For the incorporation of more realistic matter models see e.g. Burtscher & LeFloch (2014) and Burtscher (2020). Other literature may be traced back from Li & Yu (2015) and Athanasiou & Lesourd (2020).

[298] Senovilla (1998), §4, gives many examples. In the presence of a radial coordinate r as in the Schwarzschild solution, this condition is equivalent to the gradient ∇r being *timelike*, which happens for $r < 2m$.



Picture of a collapsing star, here shown to illustrate Penrose's concept of a trapped surface. Far from the collapsing matter, the local lightcones are as in Minkowski space-time, with $\theta > 0$ *and* $\underline{\theta} < 0$*, so that the area of a surface (represented by a circle in this picture) grows along L and shrinks along* $\underline{L}$*. While* $\underline{\theta} < 0$ *everywhere,* θ *vanishes at* $r = 2m$ *and then changes sign to* $\theta < 0$ *for* $0 < r < 2m$*, so that the circles shrink in both directions and the lightcones become thinner. The key point is that* $\theta < 0$ *indicates that the originally* outgoing *light-rays along L now bend* inwards *and cannot escape, justifying the relevance to black holes. Taken from Penrose (1969), Figure 2.*

- We now assume that the lightlike vector fields L and $\underline{L}$ are normalized such that

$$\nabla_L L = 0; \qquad \nabla_{\underline{L}} \underline{L} = 0. \tag{6.94}$$

The lightlike version of the expression (6.17) is then given by

$$\nabla_\mu L_\nu = k_{\mu\nu} - \tfrac{1}{2}(L_\mu \underline{L}^\rho \nabla_\rho L_\nu + L_\nu \underline{L}^\rho \nabla_\mu L_\rho), \tag{6.95}$$

which is easily verified by computing all contractions with L, $\underline{L}$, and e_i, using (6.88), (6.87), and (6.68). Another useful result,[299] still assuming (6.94), is

$$\theta = \nabla_\mu L^\mu; \qquad \underline{\theta} = \nabla_\mu \underline{L}^\mu. \tag{6.96}$$

To see this, one again uses (6.88), (6.87), and (6.68). For example, we compute

$$\begin{aligned} \theta &= h^{\mu\nu} h^\rho_\mu h^\sigma_\nu \nabla_\rho L_\sigma = h^{\rho\sigma} \nabla_\rho L_\sigma = (g^{\rho\sigma} + \tfrac{1}{2}(\underline{L}^\rho L^\sigma + L^\rho \underline{L}^\sigma)) \nabla_\rho L_\sigma \\ &= \nabla_\mu L^\mu + \tfrac{1}{2}(g(L, \nabla_{\underline{L}} L) + g(\underline{L}, \nabla_L L)) = \nabla_\mu L^\mu, \end{aligned} \tag{6.97}$$

since $g(L,L) = 0$ implies $g(L, \nabla_{\underline{L}} L) = 0$ by a calculation like (6.52), and we had (6.94).

- If $\omega_{\mu\nu} = 0$, see the comment after (6.73), as well as (6.94), the same argument as in the timelike case then implies Raychaudhuri's equation along the outward directions L, viz.[300]

$$\nabla_L \theta \equiv \dot{\theta} = -\tfrac{1}{2}\theta^2 - \sigma_{\mu\nu}\sigma^{\mu\nu} - R_{\mu\nu} L^\mu L^\nu. \tag{6.98}$$

This replaces (6.27), with a similar derivation. First, as in (6.26) we obtain

$$\nabla_L (\nabla_\mu L_\nu) = -(\nabla_\mu L^\sigma) \nabla_\sigma L_\nu + R_{\nu\rho\sigma\mu} L^\sigma L^\rho \tag{6.99}$$

straight from the derivation of the Riemann tensor and the property $A = 0$. We then substitute (6.95), contract with $g^{\mu\nu}$, and substituting (6.92), with $\omega = 0$. Analogous reasoning then leads to the following null counterpart of Proposition 6.7:[301]

Proposition 6.14 *Let $S \subset M$ be a closed spacelike surface, let $x \in S$, and let $\gamma = \gamma_L^{(x)}$ be a lightlike geodesic as above. If*

$$\theta(x) < 0; \qquad R_{\mu\nu} L^\mu L^\nu \geq 0 \tag{6.100}$$

along γ, then γ has a (later) focal point $\gamma(c)$ relative to $x = \gamma(a)$, provided that γ can indeed be extended from a to c. The same statement holds for $\underline{\gamma} = \gamma_x^{(\underline{L})}$, assuming that along $\underline{\gamma}$ we have

$$\underline{\theta}(x) < 0; \qquad R_{\mu\nu} \underline{L}^\mu \underline{L}^\nu \geq 0. \tag{6.101}$$

[299] Eq. (6.96) is a quick way to verify (6.83), e.g. $\theta = (\Gamma^\mu_{\mu 0} + \Gamma^\mu_{\mu r}) = (0 + \Gamma^\theta_{\theta r} + \Gamma^\varphi_{\varphi r}) = (1/r + 1/r) = 2/r$.

[300] One also has a similar equation for the Weingarten map W associated to k, but in the absence of an orthogonal projection $T_x C \to T_x S$, where $x \in C$, one has to replace $T_x S$ by $\tilde{T}_x S := T_x C / \mathbb{R} \cdot L_x$, with associated projection $T_x C \to \tilde{T}_x S$, $X \mapsto \tilde{X}$. Defining $W_X : \tilde{T}_x S \to \tilde{T}_X S$ by $W_x(\tilde{X}) := -\widetilde{\nabla_X L}$, one can show that along γ_L one has $\dot{W} = W^2 + \tilde{R}$, where $\tilde{R}_x : \tilde{T}_x S \to \tilde{T}_X S$ is defined by $\tilde{R}(X) := \widetilde{\Omega(X,L)L}$. This also yields (6.98). See e.g. Galloway (2017).

[301] See Hawking & Ellis (1973), Propositions 4.4.4 to 4.4.6, for details.

6.4 Penrose's singularity theorem

We now come to one of the highlights of (mathematical) GR. Proposition 6.14 will lead to a contradiction with global hyperbolicity, as in Hawking's theorem, provided some and hence all Cauchy hypersurfaces Σ are non-compact. If one envisages applications to black holes in asymptotically flat space-times (to be defined), then this seems a reasonable assumption.

Theorem 6.15 *Let (M,g) be globally hyperbolic with non-compact Cauchy surface Σ. Assume:*

1. *One has $R_{\mu\nu}\dot\gamma^\mu\dot\gamma^\nu \geq 0$ along all lightlike geodesics γ* **(null curvature condition)***;*
2. *The space-time M contains a future trapped surface.*

Then (M,g) has incomplete future-directed lightlike geodesics.

Proof. The proof is based on properties of the future horismos $E^+(S) = J^+(S)\backslash I^+(S)$.

Lemma 6.16 *Let (M,g) be a globally hyperbolic space-time and let $S \subset M$ be compact.*

1. *If M has a non-compact Cauchy surface, then $E^+(S)$ is non-compact.*
2. *If: i) assumptions 1 and 2 in the theorem hold; ii) S is a trapped surface (which is compact by definition); iii) all lightlike geodesics in M are complete, then $E^+(S)$ is compact.*

The proofs of both claims rely on the following consequence of global hyperbolicity.

Lemma 6.17 *If (M,g) is globally hyperbolic and $S \subset M$ is compact, then*

$$E^\pm(S) = \partial I^\pm(S). \tag{6.102}$$

Proof. This follows from Lemma 5.29, which makes $J^\pm(S)$ closed, and then (5.148). □

Recall that a hypersurface $A \subset M$ is *achronal* iff each timelike curve intersects it at most once, cf. (5.145). Furthermore,[302] $F \subset M$ is a *future set* if $I^+(F) \subset F$ (in other words, $I^+(x) \subset F$ for all $x \in F$), in which case ∂F is called an *achronal boundary*. Clearly, $F = I^+(S)$ is a future set, and hence $\partial I^+(S)$ is an achronal boundary; see (5.146) etc. for the proof that it is indeed achronal (the proof for general F, which we do not need, is similar). Since $I^+(S)$ is open, its boundary has codimension one (i.e. is $3d$ in $4d$ space-time) where it is smooth. The following more specific result will also be important for the analysis of event horizons (cf. §10.7):

Proposition 6.18 *Achronal boundaries are locally Lipschitz topological hypersurfaces in M.*

Knowing this,[303] the most rigorous way to prove part 1 of Lemma 6.16 is to use the following:

Proposition 6.19 *Let (M,g) be a globally hyperbolic space-time. Then any* compact *achronal topological hypersurface Σ in M is a Cauchy surface.*

[302] Penrose (1972) defines future sets through $I^+(F) = F$. We have equality in $I^+(F) \subset F$ iff F is open.

[303] See Minguzzi (2019), Theorem 2.87 (iii). To define the Lipschitz condition we equip M with a complete Riemannian metric, so that it also becomes a metric space, and ask the map $\varphi : U \to V$ in (4.123) to be bi-Lipschitz (i.e. Lipschitz with Lipschitz inverse). See e.g. Naumann & Simader (2011), §2.1. The simplest examples, such as $S = \partial I^+(0)$ in $\mathbb{M}$, whose boundary is not smooth at the apex show that achronal boundaries need not be smooth.

Proof (sketch).[304] First, the definition of a hypersurface implies that it has no boundary, which implies that $J^+(\Sigma)\cup J^-(\Sigma)$ is open. If Σ is compact, then $J^\pm(\Sigma)$ is closed by Lemma 5.29 and the assumption of global hyperbolicity. Hence $J^+(\Sigma)\cup J^-(\Sigma)$ is both open and closed, and since our space-times M are connected by definition, we must have

$$J^+(\Sigma)\cup J^-(\Sigma) = M. \tag{6.103}$$

So any $x\in M$ must lie either in $J^+(\Sigma)$ or in $J^+(\Sigma)$, or on both, i.e. in Σ, which trivial case we exclude. Suppose $x\in J^+(\Sigma)$ and consider a past inextendible timelike curve c ending at x. If c does not intersect Σ, then it must stay in $J^+(\Sigma)\cap J^-(x)$. This set is compact by Lemma 5.29, so the curve is imprisoned, which is impossible by Definition 5.27 of global hyperbolicity. □

In view of Theorem 5.33.2, which excludes the possibility of M having one Cauchy surface that is compact and another that is not, Propositions 6.18 and 6.19 clearly imply Lemma 6.16.1.[305]

We now turn to the second part of Lemma 6.16. Given (6.102), the idea is to use Corollary 5.16. The lightlike geodesics ruling $\partial I^+(S)$ come from both L and $\underline{L}$; in $d=2+1$ a picture where the trapped "surface" is just a circle shows this very clearly. This obviously gives the inclusion $\partial I^+(S)\subset C\cup\underline{C}$, see (6.61) and (6.81). The key is a refinement of this inclusion:

Lemma 6.20 *For any closed spacelike surface S in a globally hyperbolic space-time (M,g), with associated null surfaces C and $\underline{C}$ defined by (6.61) and (6.81), one has*

$$E^+(S)\subset C_{\rm reg}\cup\underline{C}_{\rm reg}\subset C\cup\underline{C}\subset J^+(S), \tag{6.104}$$

where $C_{\rm reg}\subset C$ is the regular (and hence smooth) part of C, and $\underline{C}_{\rm reg}$ is defined likewise. More precisely, $C_{\rm reg}$ is defined as the subset consisting of the parts of all (fd) lightlike geodesics in C starting in S before their first focal points (if any) are encountered (and similarly for $\underline{C}_{\rm reg}$).

The first inclusion in (6.104) follows from the counterpart of Theorem 6.6 for lightlike geodesics:

Theorem 6.21 *A lightlike geodesic $\gamma:[a,b]\to M$ in C or $\underline{C}$ (with $\gamma(a)\in S$) locally maximizes the length of causal curves from $\gamma(a)$ to $\gamma(b)$ (not necessarily in C or $\underline{C}$) iff there is* no *focal point $\gamma(c)$ on γ, $a<c<b$. Therefore, if there is an intermediate focal point, then $\gamma(b)\in I^+(\gamma(a))$.*

This is a "lightlike" adaptation of Theorem 5.12, whose long proof we omit.[306] We do note that lightlike geodesic can only maximize length if all other comparable causal curves (between the given endpoints) have zero length, too. This is why focal points and the existence of (longer) timelike curves go hand in hand. See also the comment after the proof of Proposition 5.9.

[304] For complete proofs see Budic *et al.* (1978), Theorem 1, and Galloway (1985), Theorem 1 and Corollary 1.

[305] Penrose (1965) himself, followed by Hawking & Ellis (1973), §8.2, Theorem 1, used a different argument: Since M has Cauchy surface Σ and $E^+(S)$ is achronal, via the flow of a complete timelike vector field (which exists because space-times are time orientable), any $x\in E^+(S)$ projects onto a unique point of Σ. This gives a continuous injective map $\pi: E^+(S)\to\Sigma$, which is a homeomorphism onto its image $\pi(E^+(S))$ in Σ (recall that any continuous bijection from a compact space to a Hausdorff space has a continuous inverse). Since $E^+(S)$, being a boundary itself by (6.102), has no boundary, its homeomorphic image $\pi(E^+(S))$ has no boundary either. Similarly, if $E^+(S)$ is compact, then $\pi(E^+(S))$ is compact, too. In that case the *non-compact* (sub)manifold Σ would have a compact submanifold *of the same dimension* without boundary, which is impossible (Aretakis, 2013, §5.5): any $x\in\pi(E^+(S))$ would then have an open nbhd U which is also open in Σ, since $\dim(E^+(S))=\dim\pi(E^+(S))=\dim(\Sigma)=3$. Hence $\pi(E^+(S))$ would be open in Σ, but it is also closed since it is compact. Since Σ is connected (by Proposition 5.33.1) this implies $\pi(E^+(S))=\Sigma$, which is impossible because Σ is not compact whilst $\pi(E^+(S))$ is.

[306] See Hawking & Ellis (1973), Proposition 4.5.10 to 4.5.14, O'Neill (1983), Propositions 10.46 to 10.48, or Kriele (1999), Lemma 4.6.15, Theorem 4.6.2(iii) and Corollary 8.3.1, and Minguzzi (2019), Theorem 6.16.

If now $y \in C$ lies beyond a focal point on some lightlike geodesic $\gamma_L^{(x)}$ in C, then by Corollary 5.16 it cannot lie in $\partial I^+(S)$, since now there is a timelike curve from x to y, and likewise for $\underline{L}$ and $\underline{C}$. This gives the inclusion $E^+(S) \subset \overline{C_{\rm reg}} \cup \overline{\underline{C}_{\rm reg}}$. Now suppose that all assumptions in Lemma 6.16.2 hold. Then Proposition 6.14 applies. Each (lightlike) geodesic γ in C reaches its first focal point in finite time t_f; if γ starts at $x \in S$, then $t_f(x) \leq 2/|\theta(x)|$, for which the argument is the same as after (6.34), except that in the null Raychaudhuri equation (6.98) one has $-\frac{1}{2}\theta^2$ instead of $-\frac{1}{3}\theta^2$ as in the timelike case (6.27). Since S is compact and $\theta(x) < 0$ for all $x \in S$ by definition of a trapped surface, one has $\Theta := \inf_{x\in S}\{|\theta(x)|\} > 0$, so that by the time $t_f = 2/\Theta < \infty$ each geodesic in C has passed its first focal point. Likewise for $\underline{L}$ and $\underline{C}$, giving $\underline{\Theta} = \inf_{x\in S}\{|\underline{\theta}(x)|\} > 0$ and a time $\underline{t}_f = 2/\underline{\Theta} < \infty$ playing the same role for $\underline{C}$. It follows that

$$E^+(S) \subset \overline{C_{\rm reg}} \cup \overline{\underline{C}_{\rm reg}} \subset (\cup_{t\in[0,t_f]} S_t) \bigcup (\cup_{\underline{t}\in[0,\underline{t}_f]} \underline{S}_{\underline{t}}). \tag{6.105}$$

By (6.102), this makes $E^+(S)$ a closed subset of a compact set, so that it is compact, which proves Lemma 6.16.2. Given the assumptions of Theorem 6.4, the only way out of the contradiction between compactness and non-compactness of $E^+(S)$ is to invalidate the invocation of Proposition 6.14 by using its proviso 'provided that γ can indeed be extended from a to c', which would be guaranteed by lightlike geodesics completeness. So this cannot be the case; the proof shows that at least one incomplete lightlike geodesic emanates from the trapped surface S. □

Neither Hawking's nor Penrose's singularity theorem proves the existence of a singularity in the sense of Definition 6.1.3. These theorems only show causal geodesic incompleteness and as such they are better called *incompleteness theorems*. The "singularity" theorems were inspired by intuition from the big bang and from black holes, as described by the time-honoured Schwarzschild and FLRW solutions (6.1) - (6.3), where some quantity defined via the metric becomes singular (that is, infinite or zero). However, running ahead of chapters 9 and 10, consider the Kerr solution (9.110) for $0 < |a| < m$ or even more simply the Reissner–Nordström solution (9.88) - (9.89) for $0 < |e| < m$, and suppose we do not look at the maximally extended solutions but rather at the maximal globally hyperbolic development (MGHD) of a typical maximal spacelike (achronal) hypersurface on which the initial data induced by the global solutions are given (see §7.6). The picture then changes completely: the ensuing space-times still satisfy the assumptions of Penrose's theorem, but they are not singular in any metric sense, because the singularities lie behind the Cauchy horizon of the initial data hypersurface. In this (PDE) picture, causal geodesic incompleteness rather means that the space-times in questions are *extendible*. The maximal (analytic) extensions *are* singular in a metric sense, and in addition they fail to be globally hyperbolic (whereas any MGHD is globally hyperbolic by construction). These properties are related to each other; see §10.4 in connection with (strong) cosmic censorship.

Penrose's theorem is the mother of all singularity theorems in GR and, with Hawking's (which is perhaps the father), also the cleanest one. Apart from the introduction of topological methods in GR, which was new at the time, among its main achievements one should already count a key *definition* Penrose introduced in GR, namely that of a trapped surface.

Nonetheless, there is room for weakening the assumptions in Penrose's theorem.[307] The most cited way of doing this is the combined ***Hawking–Penrose singularity theorem***:[308]

[307] As well as in Hawking's, where, as already noticed by Hawking (1967) himself, global hyperbolicity may be replaced by strong causality, in which case the Cauchy surface in its proof may be replaced by a partial one.

[308] See Hawking & Penrose (1970), Hawking & Ellis (1973), §8.2, Theorem 2, or Senovilla (1998), Theorem 5.6.

Theorem 6.22 *Let (M,g) be a chronological space-time (i.e. M contains no closed timelike curves). If $R_{\mu\nu}u^\mu u^\nu \geq 0$ for every causal vector u, and on top of that the* **genericity condition**

$$\dot\gamma_{[\alpha}R_{\beta]\gamma\delta[\rho}\dot\gamma_{\sigma]}\dot\gamma^\gamma\dot\gamma^\delta \neq 0 \tag{6.106}$$

holds in at least one point of every causal geodesic γ, and at least one of the following is present:

1. *A compact edgeless achronal set;*
2. *A closed trapped surface;*
3. *A point $x \in M$ such that the lightlike geodesics from x are focused and reconverge,*

then the space-time in question is causally geodesically incomplete.

The main achievement of this theorem is that global hyperbolicity has been weakened, and that the assumptions in Hawking's and Penrose's theorems are somehow combined. But the price is high: the assumption (6.106) is contrived and purpose-driven, and in addition the theorem does not so much *strengthen* as *weaken* Penrose's theorem:[309] because of the choice menu in its assumptions, in the case of say a black hole in an expanding universe the theorem may point towards the big bang singularity whilst saying nothing about the black hole singularity, or *vice versa*; whereas the separate theorems of Hawking and Penrose would identify both. Instead, a real and useful strengthening of Penrose's theorem is given by ***Minguzzi's singularity theorem***:

Theorem 6.23 *Let a space-time (M,g) satisfy assumptions 1 and 2 in Theorem 6.15, and also:*

1. *$I^+(x) \subset I^+(y)$ implies $I^-(y) \subset I^-(x)$ (i.e., (M,g) is* **past reflecting***);*
2. *M does* not *contain a compact spacelike hypersurface.*

Then (M,g) has incomplete future-directed lightlike geodesics.

Condition 1 weakens Penrose's assumption of global hyperbolicity. In view of Proposition 6.19, condition 2 weakens his assumption of M of containing a *non-compact* Cauchy surface. These conditions were inspired by black hole evaporation. Other assumptions one may weaken include:

- the *pointwise* energy/curvature conditions, which can be replaced by *averages*;[310]
- the presence of a trapped surface, which can be replaced by an ***outer trapped surface***;[311]
- the assumption that the space-time is chronological (which e.g. a Kerr black hole is not);[312]
- the regularity of the metric, so far tacitly assumed smooth.[313]

[309] This point was made by Minguzzi (2020), which is also the source of Theorem 6.23 below.

[310] See Fewster & Galloway (2011), Fewster & Kontou (2020), and Freivogel, Kontou, & Krommydas (2020).

[311] These things are defined in §10.11. Briefly, a (marginally) outer trapped surfaces has $\theta < 0$ ($\theta = 0$), irrespective of the sign of $\underline{\theta}$, where L is outward pointing (provided this can be defined). Outer trapped surfaces appear in the ***topological singularity theorem*** of Gannon (1975) and Lee (1976), in which assumption 2 in Theorem 6.15 is replaced by the mere assumption that Σ is non-simply connected. See also Galloway (2017), Theorem 3.3. The proof constructs an outer trapped surface in the universal cover $\tilde\Sigma$. Eichmair, Galloway, & Pollack (2013) showed that, assuming the genericity condition (6.106), condition 2 in Theorem 6.15 may be replaced by the presence of a marginally outer trapped surface in the Cauchy surface Σ. In a variation on this result, Chruściel & Galloway (2014) replaced (6.106) by assuming that the second fundamental form k defined in (6.77) is not identically zero.

[312] See Lesourd (2018). Another interesting (cosmological) singularity theorem of his is in Lesourd (2019).

[313] See Graf *et al.* (2018) and references therein. One can go down to $C^{1,1}$ (i.e. derivatives are locally Lipschitz).

7 The Einstein equations

As noticed independently by Einstein and Hilbert in 1916, the Einstein equations

$$R_{\mu\nu} - \tfrac{1}{2} g_{\mu\nu} R = 8\pi T_{\mu\nu}, \tag{7.1}$$

whose left-hand side we now understand, and whose right-hand side will be explained in §7.3, can be derived from a variational principle. The geometrical quantity to be extremized in order to obtain the left-hand side is the ***Einstein–Hilbert action*** for the gravitational field, given by

$$S_G(g) := \int_V d^4x \sqrt{-g(x)} R(x), \tag{7.2}$$

where $R = g^{\mu\nu} R_{\mu\nu}$ is the Ricci scalar. To be on the safe side in so far as convergence of integrals is concerned, $V \subset M$ is a compact region in space-time M with open interior (or, equivalently, an open region with compact closure–this does not matter with respect to a measure like d^4x), and $g \equiv \det(g)$ is the determinant of the matrix $g_{\mu\nu}$ (in any basis), cf. (3.14) - (3.15).

Eq. (7.2) will later be supplemented by boundary terms, which are needed in case things on ∂V are not under control; for the moment we omit these (life without them is difficult enough!).

7.1 Integration on manifolds

To make sense of (7.2) and its variation we need some integration theory, for which we assume some familiarity with the calculus of differential forms.[314] For simplicity we assume that M is ***orientable***, which means that there is an atlas (within the equivalence class of atlases defining the manifold, cf. §2.1) for which all transition functions $\varphi_\beta \circ \varphi_\alpha^{-1}$ have positive Jacobian. An ***orientation*** of an orientable manifold is the equivalence class of an atlas satisfying this condition. It can be shown that M is orientable iff it admits a nowhere vanishing n-form $\omega \in \Omega^n(M)$. Such an ω defines an orientation: one only accepts charts φ whose coordinates $(x^1, \ldots, x^n)$ satisfy

$$\omega(\partial_1, \ldots, \partial_n) > 0. \tag{7.3}$$

In the presence of a metric we normalize ω (which can be multiplied by an arbitrary smooth strictly positive function) such that in some, and hence in all coordinates one has, equivalently,

$$\begin{aligned} \omega(\partial_1, \ldots, \partial_n) &= \sqrt{|g|}; \\ \omega_x &= \sqrt{|g(x)|}\, dx^1 \wedge \cdots \wedge dx^n. \end{aligned} \tag{7.4}$$

[314] See e.g. Choquet-Bruhat & DeWitt-Morette (1982), chapter IV (a book I devoured as a student). Briefly, if $\dim(M) = n$ and $0 \leq p \leq n$, a ***p-form*** on M is a totally antisymmetric element of $\mathfrak{X}^{(p,0)}(M)$, cf. §2.5. These form a $C^\infty(M)$-submodule of $\mathfrak{X}^{(p,0)}(M)$, called $\Omega^p(M)$ or $\Lambda^p(M)$. One has multilinear maps $\wedge : \Omega^p(M) \times \Omega^q(M) \to \Omega^{p+q}(M)$ defined by concatenation followed by total antisymmetrization, called ***exterior multiplication***, as well as linear maps $d : \Omega^p(M) \to \Omega^{p+1}(M)$, called the ***exterior derivative***, which are uniquely characterized by the properties: i) eq. (2.56) at $p = 0$; ii) $d(\alpha \wedge \beta) = (d\alpha) \wedge \beta + (-1)^p \alpha \wedge d\beta$, where $\alpha \in \Omega^p(M)$; iii) $d^2 = 0$; iv) *locality*, in the sense that if $\alpha = \beta$ on some $U \in \mathscr{O}(M)$, then $d\alpha = d\beta$ on U. In coordinates one then has $(d\alpha)_{\mu_1 \cdots \mu_{p+1}} = \partial_{[\mu_1} \alpha_{\mu_2 \cdots \mu_{p+1}]}$. It follows that $\dim(\Omega^n_x(M)) = 1$ for all $x \in M$, with basis $dx^1 \wedge \cdots \wedge dx^n$. Finally, each vector field $X \in \mathfrak{X}(M)$ defines ***insertion*** maps $i_X : \Omega^p(M) \to \Omega^{p-1}(M)$ that are uniquely characterized by the following properties: i) $i_X = 0$ on $\Omega^0(M) \equiv C^\infty(M)$; ii) $i_X \theta = \theta(X)$ for $\theta \in \Omega^1(M) \equiv \Omega(M)$; iii) $i_X(\alpha \wedge \beta) = (i_X \alpha) \wedge \beta + (-1)^p \alpha \wedge i_X \beta$, where again $\alpha \in \Omega^p(M)$. In coordinates, one has $(i_X \alpha)_{\mu_2 \cdots \mu_p} = X^{\mu_1} \alpha_{\mu_1 \mu_2 \cdots \mu_p}$.

This expression is indeed well defined in that ω keeps this form under coordinate transformations. To see this, we use the change of coordinates formula (2.77) for the metric, i.e.,

$$g_{\mu'\nu'}(y) = \frac{\partial x^\mu}{\partial y^{\mu'}} \frac{\partial x^\nu}{\partial y^{\nu'}} g_{\mu\nu}(x), \tag{7.5}$$

in which we write y for x_β and x for x_α, so that $y = y(x)$, and as a matrix we have

$$\frac{\partial x^\mu}{\partial y^{\mu'}} \equiv \left(\frac{\partial y^{\mu'}}{\partial x^\mu} \right)^{-1}. \tag{7.6}$$

Then

$$g(y) = g(x) \left(\det \left(\frac{\partial x^\mu}{\partial y^{\mu'}} \right) \right)^2 = g(x) \left(\det \left(\frac{\partial y^{\mu'}}{\partial x^\mu} \right) \right)^{-2}. \tag{7.7}$$

This gives coordinate-independence of (7.4), as well as the (equivalent) property

$$d^4y \sqrt{|g(y)|} = d^4x \left| \det \left(\frac{\partial y^{\mu'}}{\partial x^\mu} \right) \right| \sqrt{|g(x)|} \left| \det \left(\frac{\partial y^{\mu'}}{\partial x^\mu} \right) \right|^{-1} = d^4x \sqrt{|g(x)|}. \tag{7.8}$$

The following formula then either defines the left-hand side or gives a formula for it:

$$\int_V f\omega = \int_V d^n x \sqrt{|g(x)|} f(x), \tag{7.9}$$

for any $f \in C_c^\infty(M)$, where ω is defined by (7.4), and the right-hand side should be written as a sum over various coordinate patches using a partition of unity. As we saw, the expression

$$d^4x \sqrt{|g(x)|} \tag{7.10}$$

is invariant under coordinate transformations and hence defines a "geometric" volume element.

We will encounter boundary terms. First, the ***divergence*** of a vector field X is defined as

$$\nabla \cdot X = \nabla_\mu X^\mu. \tag{7.11}$$

From now on we assume Lorentzian signature. In any coordinate system we then have

$$\partial_\mu \sqrt{-g} = \sqrt{-g} \Gamma^\rho_{\mu\rho}. \tag{7.12}$$

Indeed, since the first term in (4.15) cancels the last if $\nu = \rho$, we have

$$\Gamma^\rho_{\mu\rho} = \tfrac{1}{2} g^{\rho\sigma} g_{\rho\sigma,\mu}. \tag{7.13}$$

The right-hand side can be computed by diagonalizing the symmetric invertible matrix $(g_{\rho\sigma})$, yielding nonzero eigenvalues $(\lambda_0, \ldots, \lambda_3)$. Realizing that $(g^{\rho\sigma})$ is the inverse of $(g_{\rho\sigma})$ gives

$$g^{\rho\sigma} g_{\rho\sigma,\mu} = \frac{\partial_\mu \lambda_0}{\lambda_0} + \cdots + \frac{\partial_\mu \lambda_3}{\lambda_3}. \tag{7.14}$$

Eq. (7.12) then follows from the fact that we also have

$$2 \frac{\partial_\mu \sqrt{-g}}{\sqrt{-g}} = g^{-1} \partial_\mu g = \frac{\partial_\mu (\lambda_0 \cdots \lambda_3)}{\lambda_0 \cdots \lambda_3} = \frac{\partial_\mu \lambda_0}{\lambda_0} + \cdots + \frac{\partial_\mu \lambda_3}{\lambda_3}. \tag{7.15}$$

For later use (see §7.5) we also put on record an identity that follows from (7.12), viz.

$$\partial_\nu(\sqrt{-g}g^{\mu\nu}) = \sqrt{-g}\,g^{\rho\sigma}\Gamma^\mu_{\rho\sigma}. \tag{7.16}$$

Returning to our theme of boundary terms, eq. (7.12) implies

$$\sqrt{-g}\,\nabla\cdot X = \partial_\mu(\sqrt{-g}X^\mu), \tag{7.17}$$

and hence, by Stokes's theorem (= divergence theorem = Gauss's theorem),

$$\int_V d^4x\sqrt{-g(x)}\,\nabla\cdot X(x) = \int_{\partial V} d^3\vec{\sigma}\cdot X = \int_{\partial V} d^3\sigma^\mu X_\mu, \tag{7.18}$$

where ∂V is the boundary of V and $d^3\vec{\sigma}$ is the (outward) normal volume element of ∂V. If we use local coordinates (y^1,y^2,y^3) on ∂V, and $\tilde{g} = i^*g$ is the induced metric on ∂V (where $i:\partial V\hookrightarrow V$ is the inclusion, so that $\tilde{g}(X,Y) = g(X,Y)$ for X,Y tangent to ∂V), we have

$$d^3\vec{\sigma} = d^3y\sqrt{|\det(\tilde{g})|}\,\vec{N}, \tag{7.19}$$

where $\vec{N}$ is the outward normal to ∂V (here assumed to be non-null, so that $\det(\tilde{g})\neq 0$).
The royal (i.e. geometric) path to (7.18) is to note that (7.17) takes the abstract form

$$\mathscr{L}_X\omega = \omega\,\nabla\cdot X, \tag{7.20}$$

where the volume form ω is given by (7.4), and then use ***Cartan's formula***

$$\mathscr{L}_X\alpha = d(i_X\alpha) + i(d\alpha) \tag{7.21}$$

for the Lie derivative of any p-form $\alpha\in\Omega^p(M)$, $p>0$, where $X\in\mathfrak{X}(M)$. Since $\omega\in\Omega^n(M)$ we must have $d\omega = 0$, so that Cartan's formula for the volume form is

$$\mathscr{L}_X\omega = d(i_X\omega). \tag{7.22}$$

With (7.20), this gives $\omega\,\nabla\cdot X = d(i_X\omega)$. The abstract version of ***Stokes's theorem*** reads

$$\int_V d\alpha = \int_{\partial V}\alpha, \tag{7.23}$$

for any $\alpha\in\Omega^n(M)$. Taking $\alpha = i_X\omega$ and hence $d\alpha = \mathscr{L}_X\omega$ gives (7.18) geometrically:

$$\int_V \omega\,\nabla\cdot X = \int_{\partial V} i_X\omega. \tag{7.24}$$

Moreover, the form in (7.19) is $\sigma = i_{\vec{N}}\omega$. Using (7.21) twice, as well as (7.20), we obtain

$$\mathscr{L}_{\vec{N}}\sigma = \mathscr{L}_{\vec{N}}i_{\vec{N}}\omega = i_{\vec{N}}d(i_{\vec{N}}\omega) = i_{\vec{N}}\mathscr{L}_{\vec{N}}\omega = i_{\vec{N}}\omega\nabla\cdot\vec{N} = \sigma\nabla\cdot\vec{N}. \tag{7.25}$$

This formula will not be used in what follows, but it was used in §6.1 and hence needed proof.

7.2 Variation of the Einstein–Hilbert action

In order to set up the variational calculus around (7.2), as in the geodesic case (§3.2) we now consider a family of metrics g_s, and compute $dS_G(g_s)/ds$. This requires some preparation.

1. Each of the three terms in the integrand $\sqrt{-g}\,g^{\mu\nu}R_{\mu\nu}$ in (7.2) depends on the metric $g_{\mu\nu}$ and hence has to be varied. The variation of the Ricci tensor seems the most complicated case, but surprisingly it contributes a divergence term and hence makes no contribution to the Einstein equations (7.1). This is surprising, since definitions (4.14) and (4.109) give

$$R_{\mu\nu} = \Gamma^\rho_{\mu\nu,\rho} - \Gamma^\rho_{\mu\rho,\nu} + \Gamma^\rho_{\rho\sigma}\Gamma^\sigma_{\nu\mu} - \Gamma^\rho_{\nu\sigma}\Gamma^\sigma_{\rho\mu}, \tag{7.26}$$

whose first two terms contain second-order derivaties of $g_{\mu\nu}$. Their variation would therefore in principle be expected to give a fourth-order PDE, but this does not happen.[315]

2. Writing $\delta F(g) = dF(g_s)/ds_{|s=0}$ and $d(g_s)_{\mu\nu}/ds_{|s=0} = \delta g_{\mu\nu} \equiv d_{\mu\nu}$, we claim that

$$g^{\mu\nu}\delta R_{\mu\nu} = \nabla\cdot X; \tag{7.27}$$

$$X^\mu = \nabla^\nu d^\mu_\nu - \nabla^\mu d^\nu_\nu, \tag{7.28}$$

where indices are always raised and lowered with the metric $g = g_{s=0}$. In this respect the notation $\delta g^{\mu\nu}$ is ambiguous, as it could mean either $(\delta g)^{\mu\nu} = g^{\mu\rho}g^{\nu\sigma}\delta g_{\rho\sigma}$, or

$$\delta g^{\mu\nu} \equiv \delta(g^{\mu\nu}) = -g^{\mu\rho}g^{\nu\sigma}d_{\rho\sigma}, \tag{7.29}$$

which is what we will take it to mean. Then (7.29) follows from $g^{\mu\nu}g_{\nu\rho} = \delta^\mu_\rho$, and hence

$$0 = \delta(g^{\mu\nu}g_{\nu\rho}) = (\delta g^{\mu\nu})g_{\nu\rho} + g^{\mu\nu}d_{\nu\rho}. \tag{7.30}$$

The key step in the proof of (7.27) - (7.28) is the relation

$$\delta\Gamma^\rho_{\mu\nu} = \tfrac{1}{2}(\nabla_\mu d^\rho_\nu + \nabla_\nu d^\rho_\mu - \nabla^\rho d_{\mu\nu}) = \tfrac{1}{2}g^{\rho\sigma}(\nabla_\mu d_{\sigma\nu} + \nabla_\nu d_{\sigma\mu} - \nabla_\sigma d_{\mu\nu}). \tag{7.31}$$

This can be shown by a lengthy computation, but also by the following instructive trick.

   (a) First note that although the coefficients $\Gamma^\rho_{\mu\nu}$ do not form the components of a tensor, their variation $\delta\Gamma^\rho_{\mu\nu}$ does. This is true far more generally: if ∇ and $\tilde\nabla$ are connections on a vector bundle E, then $(\nabla_X - \tilde\nabla_X)s$ is $C^\infty(M)$-linear in $s \in \Gamma(E)$ (unlike $\nabla_X s$ and $\tilde\nabla_X s$), since the spoiler $(Xf)s$ in the Leibniz rule (3.56) drops out of the difference. As a case in point, let ∇ be the Levi-Civita connection for a given metric g and let $\tilde\nabla$ be the Levi-Civita connection for some other metric $\tilde g$. We then have a tensor $\hat\Gamma \in \mathfrak{X}^{(2,1)}(M)$, defined by $\hat\Gamma(X,Y,\theta) = \theta(\nabla_X Y - \tilde\nabla_X Y)$, whose connection coefficients are $\Gamma^\rho_{\mu\nu} - \tilde\Gamma^\rho_{\mu\nu}$, cf. (3.37). In particular, we make take $\tilde g = g_s$, and since

$$\delta\Gamma^\rho_{\mu\nu}(g) = \lim_{s\to 0}(\Gamma^\rho_{\mu\nu}(g_s) - \Gamma^\rho_{\mu\nu}(g))/s, \tag{7.32}$$

we may conclude that the coefficients $\delta\Gamma^\rho_{\mu\nu}$ form the components of a tensor $\delta\Gamma$.

[315] ***Lovelock's Theorem*** (Lovelock, 1971; Navarro & Navarro, 2010) states that in $d = 4$ the Einstein–Hilbert action (7.2) is the *only* possible geometric quantity giving rise to second-order PDE in the $g_{\mu\nu}$, except for adding a (cosmological) constant $\Lambda = -\lambda$ to the Ricci scalar R. See Anderson (1981) for an extension to matter couplings.

(b) Let σ and τ be tensors of the same type, say $(1,1)$ Then $\sigma=\tau$ is true iff for each $x\in M$ one has $\sigma^\nu_\mu(x)=\tau^\nu_\mu(x)$ in just *one* specific coordinate system (x^μ) defined on some nbhd U of x, which system may even depend on x (like GNC). For in that case we have $\sigma_x(\partial_\mu,dx^\nu)=\tau_x(\partial_\mu,dx^\nu)$, and so, by $C^\infty(M)$-linearity of σ and τ in its arguments, $\sigma(X,\theta)=\tau(X,\theta)$, where we write $X=X^\mu\partial_\mu$ and $\theta=\theta_\nu dx^\nu$ as usual, for some $X^\mu\in C^\infty(U)$ and $\theta_\nu\in C^\infty(U)$. And similarly for tensors of any type (k,l).

(c) It therefore suffices to verify (7.31) in geodesic normal coordinates, where at $x=x_0$ we have $\nabla=\partial$, cf. (5.38). In GNC one does not even need (7.29), since $\delta g^{\rho\sigma}$ in (4.15) multiplies terms that vanish at x_0, and hence (7.31) is almost trivial.

Similarly, noting that in GNC the variation $\delta R_{\mu\nu}$ only employs the first two terms in (7.26), in which $\delta(\Gamma^\rho_{\mu\nu,\rho})=\partial_\rho\delta\Gamma^\rho_{\mu\nu}$ (etc.) can be computed from (7.31), one obtains

$$\delta R_{\mu\nu}=\tfrac{1}{2}(\nabla_\rho\nabla_\mu d^\rho_\nu+\nabla_\rho\nabla_\nu d^\rho_\mu-\nabla_\mu\nabla_\nu d^\rho_\rho-\nabla^\rho\nabla_\rho d_{\mu\nu}),\tag{7.33}$$

where we note that the third term is symmetric in μ and ν because of (4.13) and (4.37). Contraction with $g^{\mu\nu}$ then makes the first two terms identical to each other, and similarly, the last two. This immediately leads to (7.27) - (7.28).

3. The computation of $\delta\sqrt{-g}$ is based on the relation $\partial g/\partial g_{\mu\nu}=g^{\mu\nu}g$,[316] which implies

$$\delta\sqrt{-g}=\frac{\partial\sqrt{-g}}{\partial g_{\mu\nu}}d_{\mu\nu}=-\frac{1}{2\sqrt{-g}}\frac{\partial g}{\partial g_{\mu\nu}}d_{\mu\nu}=\tfrac{1}{2}\sqrt{-g}\,g^{\mu\nu}d_{\mu\nu}.\tag{7.34}$$

4. Since we already know $\delta g_{\mu\nu}$ from (7.29), we are finally in a position to compute:

$$\begin{aligned}S'_G(g)&=\frac{dS_G(g_s)}{ds}(s=0)=\int_V d^4x\,\delta(\sqrt{-g}g^{\mu\nu}R_{\mu\nu})\\&=\int_V d^4x\,[(\delta\sqrt{-g})g^{\mu\nu}R_{\mu\nu}+\sqrt{-g}\,(\delta g^{\mu\nu})R_{\mu\nu}+\sqrt{-g}\,g^{\mu\nu}\delta R_{\mu\nu}]\\&=\int_V d^4x\sqrt{-g}\,(\tfrac{1}{2}g^{\mu\nu}R-R^{\mu\nu})d_{\mu\nu}+\int_{\partial V}d^3\vec{\sigma}^\mu(\nabla^\nu d_{\mu\nu}-\nabla_\mu d^\nu_\nu).\end{aligned}\tag{7.35}$$

5. Now a delicate point is that although by definition of the variational principle $d_{\mu\nu}$ vanishes on the boundary ∂V, this need not be the case for its (covariant) derivatives $\nabla^\nu d_{\mu\nu}$ etc. To cancel the problematic boundary term in (7.35) one needs to add a boundary term $S_B(g)$ to the Einstein–Hilbert action (7.2), giving a gravitational action $S=S_G+S_B$, where

$$S_B(g):=2\int_{\partial V}d^3\vec{\sigma}\cdot\vec{N}\,(\mathrm{Tr}\,(\tilde{k}^0)-\mathrm{Tr}\,(\tilde{k}))=2\varepsilon\int_{\partial V}d^3y\sqrt{|\det(\tilde{g})|}\,(\mathrm{Tr}\,(\tilde{k}^0)-\mathrm{Tr}\,(\tilde{k})).\tag{7.36}$$

Here $\varepsilon=g(\vec{N},\vec{N})$ equals $\varepsilon=1$ if ∂V is timelike and $\varepsilon=-1$ if ∂V is spacelike; in a 3+1 split, where V typically looks like the bulk part of a cylinder, ∂V consists of two timelike components that bound V from above and from below, as well as a single spacelike part (cf. §8.7). Furthermore, $\tilde{k}$ is the trace of the extrinsic curvature of the embedding $\partial V\hookrightarrow V$. The extrinsic curvature, studied in detail in §4.7, is defined by

$$\tilde{k}(X,Y):=-g(\nabla_X N,Y),\tag{7.37}$$

[316] This follows from linear algebra: $\partial g/\partial g_{\mu\nu}=m^{\mu\nu}$, i.e. the minor = cofactor of $g_{\mu\nu}$, and $g^{\mu\nu}=m^{\nu\mu}/g$.

where we dropped the arrow on the normal $\vec{N}$, and X, Y are tangent to ∂V; this form turns out to be a symmetric tensor $\tilde{k} \in \mathfrak{X}^{(2,0)}\partial V$ on ∂V, like the induced metric $\tilde{g}$, cf. (4.146). Nonetheless, there is a convenient space-time calculus for both $\tilde{g}$ and $\tilde{k}$, defined via

$$\tilde{g}_{\mu\nu} := g_{\mu\nu} - \varepsilon N_\mu N_\nu; \tag{7.38}$$

$$\tilde{k}_{\mu\nu} := -\tilde{g}^\rho_\mu \tilde{g}^\sigma_\nu \nabla_\rho N_\sigma, \tag{7.39}$$

whose indices are raised and lowered with g. However, since $N_\mu N_\nu \tilde{k}_{\mu\nu} = 0$, the trace is

$$\mathrm{Tr}\,(\tilde{k}) = \tilde{g}^{\mu\nu}(\Gamma^\rho_{\mu\nu} N_\rho - \partial_\mu N_\nu). \tag{7.40}$$

Finally, $\tilde{k}^0$ in (7.36) is the extrinsic curvature of the embedding $\partial V \hookrightarrow \mathbb{M}$, where $\mathbb{M}$ is Minkowski space-time (its trace is taken with respect to the Minkowski metric). This term is necessary for for (7.36) to converge if ∂V stretches out to spatial infinity.

In computing the variation of $S_B(g)$, it simplifies matters greatly that $d_{\mu\nu}$ vanishes on ∂V, as do all its derivatives along ∂V, so that only its derivatives along N need to be taken into account. For example, $\delta \det(\tilde{g})$ vanishes on ∂V, and for $\delta \mathrm{Tr}\,(\tilde{k})$ on ∂V we find

$$\delta \mathrm{Tr}\,(\tilde{k}) = \tilde{g}^{\mu\nu} N_\rho \delta \Gamma^\rho_{\mu\nu} = -\tfrac{1}{2} \tilde{g}^{\mu\nu} N^\rho \partial_\rho d_{\mu\nu}, \tag{7.41}$$

so that

$$\delta S_B(g) = \varepsilon \int_{\partial V} d^3y \sqrt{|\det(\tilde{g})|}\, \tilde{g}^{\mu\nu} N^\rho \partial_\rho d_{\mu\nu}. \tag{7.42}$$

On the other hand, on ∂V where $d_{\mu\nu} = 0$, we have, for the boundary term in (7.35),

$$\begin{aligned} N^\mu(\nabla^\nu d_{\mu\nu} - \nabla_\mu d^\nu_\nu) &= N^\mu g^{\alpha\beta}(\partial_\alpha d_{\mu\beta} - \partial_\mu d_{\alpha\beta}) \\ &= N^\mu(\tilde{g}^{\alpha\beta} + \varepsilon N^\alpha N^\beta)(\partial_\alpha d_{\mu\beta} - \partial_\mu d_{\alpha\beta}) \\ &= -\tilde{g}^{\alpha\beta}\partial_\mu d_{\alpha\beta}, \end{aligned} \tag{7.43}$$

since $\tilde{g}^{\alpha\beta}\partial_\alpha d_{\mu\beta} = 0$ on ∂V and $N^\mu N^\alpha N^\beta(\partial_\alpha d_{\mu\beta} - \partial_\mu d_{\alpha\beta}) = 0$ identically. Hence

$$\int_{\partial V} d^3\vec{\sigma}^\mu(\nabla^\nu d_{\mu\nu} - \nabla_\mu d^\nu_\nu) = -\varepsilon \int_{\partial V} d^3y \sqrt{|\det(\tilde{g})|}\, \tilde{g}^{\alpha\beta} N^\mu \partial_\mu d_{\alpha\beta}, \tag{7.44}$$

where we used (7.19), so that the last term in (7.35) cancels (7.42), as intended.

6. In view of these computations, we obtain for the variation of $S(g) = S_G(g) + S_B(g)$:

$$\delta S(g) = \int_V d^4x \sqrt{-g}\,(R_{\mu\nu} - \tfrac{1}{2} g_{\mu\nu} R)\delta g^{\mu\nu}, \tag{7.45}$$

where we used (7.29). If there were no matter in the universe, then requiring $S'_G(g) = 0$ for arbitrary variations $d_{\mu\nu}$ (or $\delta g^{\mu\nu}$) therefore gives the ***vacuum Einstein equations***

$$R_{\mu\nu} - \tfrac{1}{2} g_{\mu\nu} R = 0. \tag{7.46}$$

It was a fact of great importance to Einstein that the gravitational action (7.2) is, as he called it, *generally covariant*, i.e., invariant under arbitrary coordinate transformations. See also §1.10.

We would now rather say that $S_G(g)$ is invariant under (orientation-preserving) diffeomorphisms. This has a very interesting consequence.[317] Consider variations of the metric that take the form

$$g_s = \varphi_s^* g, \tag{7.47}$$

where φ_s is a one-parameter group of diffeomorphisms of M that preserve V, arising as the flow of a vector field $X \in \mathfrak{X}(M)$ having compact support within V (in which case X is complete). As a special case of (2.93), for the above variations (7.47) we have

$$\frac{dg_s}{ds}(s=0) = \mathscr{L}_X g, \tag{7.48}$$

and hence, using (3.73).

$$d_{\mu\nu} = \nabla_\mu X_\nu + \nabla_\nu X_\mu. \tag{7.49}$$

Although this may seem obvious, we now explicitly show that

$$S_G(\varphi^* g) = S_G(g). \tag{7.50}$$

Indeed, starting from $S_G(g) = \int_V \omega_g R_g$, where we have now explicitly indicated the g-dependence of ω and R, we obtain $\varphi^* \omega_g = \omega_{\varphi^* g}$ and $\varphi^* R_g = R_{\varphi^* g}$, so that

$$\omega_{\varphi^* g} R_{\varphi^* g} = \varphi^* \omega_g \varphi^* R_g = \varphi_*(\omega_g R_g). \tag{7.51}$$

For any top-dimensional form $\alpha \in \Omega^n(M)$ (with compact support) one has

$$\int_V \varphi^* \alpha = \int_V \alpha, \tag{7.52}$$

so the transformed action equals

$$S_G(\varphi^* g) = \int_V \omega_{\varphi^* g} R_{\varphi^* g} = \int_V \varphi^*(\omega_g R_g) = \int_V \omega_g R_g = S_G(g). \tag{7.53}$$

Therefore, for variations of the kind (7.47) we have $S_G'(g) = 0$ for *any* metric g, that is, *whether or not g solves the vacuum Einstein equations*; the latter guarantee that $S_G'(g) = 0$ under *arbitrary* variations of g, as opposed to the special ones (7.47). Using the Einstein tensor

$$G_{\mu\nu} := R_{\mu\nu} - \tfrac{1}{2} g_{\mu\nu} R, \tag{7.54}$$

which like $R_{\mu\nu}$ and $g_{\mu\nu}$ is symmetric, from (7.50) we therefore have

$$\begin{aligned} 0 = S_G'(g) &= -\int_V d^4x \sqrt{-g}\, G^{\mu\nu}(\nabla_\mu X_\nu + \nabla_\nu X_\mu) \\ &= 2\int_V d^4x \sqrt{-g}\, (\nabla_\mu G^{\mu\nu}) X_\nu - 2\int_V d^4x \sqrt{-g}\, \nabla_\mu (G^{\mu\nu} X_\nu) \\ &= 2\int_V d^4x \sqrt{-g}\, (\nabla_\mu G^{\mu\nu}) X_\nu, \end{aligned} \tag{7.55}$$

since as in (7.45) the second term in the middle line is a boundary integral, which vanishes since X was assumed to have compact support within V. The final term must then vanish for arbitrary X. This recovers the ***(contracted) Bianchi identity***, which holds, once again, *for any metric g*:

$$\nabla_\mu G^{\mu\nu} = 0. \tag{7.56}$$

This also follows from (4.25) and (7.54). Its impact on GR will be studied in §7.5.

[317] These also follow from Noether's second theorem (cf. footnote 100), but are in fact easier to understand directly.

7.3 The energy-momentum tensor

The left-hand side of the Einstein equation (7.1) describes the geometry of space-time. The right-hand side $T_{\mu\nu}$ (times 8π), called the ***energy-momentum tensor***, describes the matter content of the universe. The first thing one infers from (7.1) is that $T \in \mathfrak{X}^{(2,0)}(M)$ has to satisfy

$$T_{\mu\nu} = T_{\nu\mu}. \tag{7.57}$$

This makes index raising unambiguous, so that we may write T^μ_ν for either $g^{\mu\rho}T_{\rho\nu}$ or $g^{\mu\rho}T_{\nu\rho}$. Relative to a swarm of observers whose four-velocities u, normalized by (6.4), comprise a (local) timelike congruence (cf. §6.1), the ***energy-momentum four-vector*** of matter is $T^\mu_\nu u^\nu$, and

$$E = T(u,u) = T_{\mu\nu}u^\mu u^\nu \tag{7.58}$$

is the (relative) ***energy density*** of the matter. Similarly, one has a (covariant) ***momentum density***

$$P^\mu = -h^\mu_\nu T^\nu_\rho u^\rho, \tag{7.59}$$

cf. (6.11), which is orthogonal to u, i.e., $g(P,u) = 0$. The fully orthogonal projection of T, viz.

$$S_{\mu\nu} = h^\rho_\mu h^\sigma_\nu T_{\rho\sigma}, \tag{7.60}$$

is the ***stress tensor*** (of the given matter): if X and Y are spacelike unit vectors orthogonal to u, then $S(X,Y)$ is the force exerted by the matter in the direction X on the spacelike unit surface element normal to Y, and *vice versa*, since $S(X,Y) = S(Y,X)$. This gives the decomposition

$$T_{\mu\nu} = S_{\mu\nu} + P_\mu u_\nu + P_\nu u_\mu + E u_\mu u_\nu. \tag{7.61}$$

Since the Einstein equations may be rewritten as

$$R_{\mu\nu} = 8\pi(T_{\mu\nu} - \tfrac{1}{2}g_{\mu\nu}T), \tag{7.62}$$

where $T = T^\mu_\mu = g^{\mu\nu}T_{\mu\nu}$ is the trace of T, it is often useful to know that, as implied by (7.61),

$$T = S - E, \tag{7.63}$$

where $S = g^{\mu\nu}S_{\mu\nu}$ is purely spatial, i.e. $S = \sum_{i=1}^3 S(e_i,e_i)$ for some o.n.b. (e_i) orthogonal to u. Assuming (7.1), the curvature condition (6.34) in Hawking's Theorem 6.4 is equivalent to

$$E \geq -S. \tag{7.64}$$

More generally, the most straightfoward ***energy conditions*** used in GR are the following:

$$T_{\mu\nu}X^\mu Y^\nu \geq 0, \quad X \sim Y \text{ causal}, \quad \textbf{\textit{(dominant energy condition = DEC)}}; \tag{7.65}$$

$$T_{\mu\nu}X^\mu X^\nu \geq 0, \quad X \text{ causal}, \quad \textbf{\textit{(weak energy condition = WEC)}}; \tag{7.66}$$

$$T_{\mu\nu}X^\mu X^\nu \geq \tfrac{1}{2}X^\mu X_\mu T, \quad X \text{ causal}, \quad \textbf{\textit{(strong energy condition = SEC)}}; \tag{7.67}$$

$$T_{\mu\nu}X^\mu X^\nu \geq 0, \quad X \text{ timelike}, \quad \textbf{\textit{(null energy condition = NEC)}}, \tag{7.68}$$

where $X \sim Y$ denotes that X and Y, both causal, should be either both fd or both pd. One has obvious implications DEC $\Rightarrow$ WEC $\Rightarrow$ NEC and SEC $\Rightarrow$ NEC, and DEC is equivalent to WEC plus the requirement that $T^\mu_\nu X^\nu$ be causal for causal X. As such, it may be strengthened by the ***strengthened dominant energy condition = SDEC***, which requires (7.66) plus the condition that $T^\mu_\nu X^\nu$ be *timelike* for *timelike* X, provided $T_{\mu\nu} \neq 0$. DEC will be used e.g. in black hole thermodynamics, cf. Proposition 10.38. Here is a completely different application of DEC:[318]

[318] See Malament (2012), Prop. 2.5.1, Hawking & Ellis, §4.3, and, in final form, Minguzzi (2015b).

Proposition 7.1 *Suppose a symmetric tensor $T_{\mu\nu}$ satisfies DEC and the* **conservation law**

$$\nabla^\mu T_{\mu\nu} = 0. \tag{7.69}$$

If $S \subset M$ is an achronal set on which $T_{\mu\nu} = 0$, then $T_{\mu\nu}$ also vanishes on $D(S)$, cf. (5.170).

If $T_{\mu\nu}$ is "the" energy-momentum tensor, then (7.69) follows either from the Bianchi identity (7.56) and Einstein's equation (7.1), or, if $T_{\mu\nu}$ can be derived from an action principle, from an argument like the one at the end of §7.2. To see SDEC in action, we mention another difficult result, making Einstein's idea that (7.1) implies geodesic motion of test particles rigorous:[319]

Proposition 7.2 *Suppose a symmetric tensor $T_{\mu\nu}$ satisfies SDEC and and* (7.69). *Let $c : I \to M$ be a curve such that $T_{\mu\nu} = 0$ outside any nbhd of $c(I)$ but $T_{\mu\nu}(c(t)) \neq 0$ for some $t \in I$. Then c can be reparametrized (if necessary) so as to become a timelike geodesic, cf.* (3.48)

The idea is that $T_{\mu\nu}$ describes a point-like "test-particle", which moves under the influence of gravity but does not act as a source. Note that the Einstein equations (7.1) are not even assumed!

A much simpler result can be derived for so-called ***dust***, with energy-momentum tensor

$$T_{\mu\nu} = \rho u_\mu u_\nu, \tag{7.70}$$

where $\rho \in C^\infty(M)$ is the mass density and u is as above, including (6.4). Eq. (7.69) gives

$$\nabla_\mu(\rho u^\mu) \cdot u + \rho \nabla_u u = 0. \tag{7.71}$$

Since $g(u, \nabla_u u) = 0$ because of (7.69), contraction with u yields two independent conditions

$$\nabla_\mu(\rho u^\mu) = 0; \qquad\qquad \nabla_u u = 0, \tag{7.72}$$

of which the first is a conservation law and the second is just the geodesic equation for u. Eq. (7.70) is a special case of the energy-momentum tensor of a ***perfect fluid***, which is given by

$$T_{\mu\nu} = (\varepsilon + p) u_\mu u_\nu + p g_{\mu\nu} = \varepsilon u_\mu u_\nu + p h_{\mu\nu}, \tag{7.73}$$

where the energy density ε is related by the pressure density p through some equation of state, such as $p = 0$ (dust, as above) or $p = \frac{1}{3}\varepsilon$ (ultrarelativistic fluid). Eq. (7.69) now gives

$$(\varepsilon + p)\nabla_\mu u^\mu + u(\varepsilon) = 0; \qquad\qquad (\varepsilon + p)\nabla_u u^\mu + h^{\mu\nu}\partial_\nu p = 0, \tag{7.74}$$

called the (relativistic) ***Euler equations***. The quantities (7.58) - (7.60) are obviously given by

$$E = \varepsilon; \qquad\qquad P = 0; \qquad\qquad S_{\mu\nu} = p h_{\mu\nu}, \tag{7.75}$$

so that $S = 3p$ and $T = 3p - \varepsilon$. The energy conditions then come down to (nontrivial exercise!):

- SEC holds iff $\varepsilon + p \geq 0$ and $\varepsilon + 3p \geq 0$;
- WEC holds iff $\varepsilon + p \geq 0$ and $\varepsilon \geq 0$;
- DEC and SDEC coincide in the case of (7.73) and both hold iff $\varepsilon \geq |p|$.

[319] This idea goes back to Einstein & Gommer (1927) and Einstein, Infeld, & Hoffman (1938). For Proposition 7.2 see Geroch & Jang (1975), as well as Geroch & Weatherall (2018) for further results. As in footnote 289, we refer to Curiel (2014a) and Martín-Moruno & Visser (2017) for more information about energy conditions.

Except for fluids,[320] most energy-momentum tensors are derived from an action principle, like the Einstein equations. The idea is that the "coupling" of gravity to matter is described by a functional $S_M(g,\varphi)$, where φ stands for all matter fields, so that, analogously to (7.45), one has

$$S'_M(g,\varphi) = -\tfrac{1}{2}\int_V d^4x\,\sqrt{-g}\,T_{\mu\nu}\delta g^{\mu\nu}, \tag{7.76}$$

where the prime has the same meaning as in §7.2 (varying the metric), or, as physicists write,[321]

$$T_{\mu\nu} = -2\frac{\delta S_M(g,\varphi)}{\delta g^{\mu\nu}}. \tag{7.77}$$

In this notation, the Einstein equation (7.1) then simply states that

$$\frac{\delta}{\delta g^{\mu\nu}}(S_G(g) + 16\pi S_M(g,\varphi)) = 0. \tag{7.78}$$

This equation for the metric $g_{\mu\nu}$ is to be supplemented with equations for the field(s), viz.[322]

$$\frac{\delta S_M(g,\varphi)}{\delta\varphi} = 0. \tag{7.79}$$

The simplest example is a ***scalar field*** $\varphi \in C^\infty(M)$, whose action functional is

$$S_M(g,\varphi) = -\tfrac{1}{2}\int_V d^4x\,\sqrt{-g}\,(g^{\mu\nu}\partial_\mu\varphi\partial_\nu\varphi + V(\varphi)) \equiv -\tfrac{1}{2}\int_V (g(\nabla\varphi,\nabla\varphi) + V(\varphi)), \tag{7.80}$$

where $V:\mathbb{R}\to\mathbb{R}$ is a "potential" (which for a free field equals $V(\varphi) = \frac{1}{2}m^2\varphi^2$). The computation (7.45), with $R_{\mu\nu}$ replaced by $\partial_\mu\varphi\partial_\nu\varphi$ (so that there isn't even a boundary term) gives

$$T_{\mu\nu} = \partial_\mu\varphi\partial_\nu\varphi - \tfrac{1}{2}g_{\mu\nu}(g(\nabla\varphi,\nabla\varphi) + V(\varphi)). \tag{7.81}$$

Another case of interest is the ***electromagnetic field*** $A \in \Omega^1(M)$, with $F = dA \in \Omega^2(M)$, or

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu = \nabla_\mu A_\nu - \nabla_\nu A_\mu, \tag{7.82}$$

where the last equality follows because ∇ is torsion-free. The (free) action is

$$S_M(g,A) = -\frac{1}{8\pi}\int_V d^4x\,\sqrt{-g}\,g^{\mu\rho}g^{\nu\sigma}F_{\mu\nu}F_{\rho\sigma} \equiv -\frac{1}{8\pi}\int_V F^2, \tag{7.83}$$

with $F^2 = F_{\mu\nu}F^{\mu\nu}$, from which a brief computation yields the energy-momentum tensor

$$T_{\mu\nu} = \frac{1}{4\pi}(g^{\rho\sigma}F_{\mu\rho}F_{\nu\sigma} - \tfrac{1}{4}g_{\mu\nu}F^2), \tag{7.84}$$

where the last term comes from the variation of $\sqrt{-g}$ and the first one comes from $\delta(g^{\mu\rho}g^{\nu\sigma})$. For later use (see §§10.9–10.10), we note that (7.84) satisfies DEC, and hence certainly NEC.

[320] Even for ideal fluids one has a (constrained) action principle due to A.H. Taub, but it is extremely contrived.

[321] In order to obtain the correct Einstein equations one is, of course, free to vary prefactors and even signs in (7.77) and (7.78), but our choice matches the convention for $T_{\mu\nu}$ in quantum field theory, with respect to which one should actually multiply Newton's constant G with the factor 16π in (7.78) and with 8π in (7.1).

[322] We might as well write these as $\delta(S_G(g) + S_M(g,\varphi))/\delta\varphi = 0$, since $S_G(g)$ is independent of φ.

7.4 Electromagnetism: gauge invariance and constraints

We start with ***electromagnetism***, since it allows us to make an important conceptual point with regard to the Einstein equations. To make this point it is enough to work in Minkowski spacetime, in which $\nabla_\mu = \partial_\mu$, $A^0 = -A_0$, $A^i = A_i$ ($i = 1,2,3$), and the wave operator (d'Alembertian)

$$\Box := -\partial_t^2 + \Delta. \tag{7.85}$$

The equation of motion for A_μ, obtained by varying A_μ in (7.83) with flat metric $g_{\mu\nu} = \eta_{\mu\nu}$, is

$$\frac{\delta S_M(g,A)}{\delta A_\mu} = \frac{\partial \mathscr{L}}{\partial A_\mu} - \partial_\nu \frac{\partial \mathscr{L}}{\partial(\partial_\nu A_\mu)} = 0. \tag{7.86}$$

For the specific action (7.83) this immediately yields

$$R_\mu = 0; \qquad R_\mu := \partial^\nu F_{\nu\mu} = \Box A_\mu - \partial_\mu(\partial_\nu A^\nu), \tag{7.87}$$

which may, more intrinsically,[323] be written in terms of the Hodge dual as $d*F = 0$ (the other half of the Maxwell equations is $dF = 0$, which however is automatic given $F = dA$). In parallel with the discussion in §7.2.7, the action (7.83) is ***gauge invariant***, in that we have $S_M(A + d\lambda) = S_M(A)$, say for all $\lambda \in C_c^\infty(V)$. Gauge invariance under $\delta A_\mu = \partial_\mu \lambda$ yields

$$0 = \int_V d^4x\, \partial_\nu F^{\nu\mu} \partial_\mu \lambda = -\int_V d^4x\, \lambda \partial_\mu \partial_\nu F^{\nu\mu} \tag{7.88}$$

for all $\lambda \in C_c^\infty(V)$, which gives the ***Bianchi identity for electromagnetism***, $\partial_\mu \partial_\nu F^{\nu\mu} = 0$, i.e.

$$\partial_\mu R^\mu = 0. \tag{7.89}$$

This is so obvious (in view of the antisymmetry of F) as to be disappointing, but it must be stressed that (7.89) is similar to (7.56) in being an *identity*, which holds irrespective of the equations of motion. See below for its thrust! Another consequence of gauge invariance is that

the equations of motion (7.87) *are simultaneously underdetermined and overdetermined*:

- They are *underdetermined* in that: if A solves (7.87), then so does $A + d\lambda$, $\lambda \in C_c^\infty(\mathbb{R}^4)$;
- They are *overdetermined* in that the initial values are constrained (i.e. cannot be arbitrary).

The first point is immediate from (7.87). For the second, we note that for $\mu = 0$ eq. (7.87) reads

$$C = 0; \qquad C := R_0 = \partial^\nu F_{\nu 0} = \partial_i F_{i0} = \Box A_0 - \partial_0(\partial_\nu A^\nu) = \Delta A_0 - \partial_0(\nabla \cdot \vec{A}). \tag{7.90}$$

This is not an evolution equation but a ***constraint*** on the initial data $A_\mu(\vec{x})$ and $\dot{A}_\mu(\vec{x})$ at $x^0 \equiv t = 0$, $\vec{x} = (x^1, x^2, x^3)$. The fact that R_0 does not contain second-order derivatives in time follows from the "Bianchi identity" (7.89), for if $\partial_t R^0$ equals some expression containing at most second-order derivatives in time, then $R^0 = -R_0$ contains at most first-order derivatives in time. Since (7.89) follows from the gauge invariance of the action that causes the underdetermination, we see that under- and overdetermination of the field A_μ are two sides of the same coin. Defining the ***electric field*** $E_i = F_{i0} = \partial_i A_0 - \partial_0 A_i$ ($i = 1,2,3$), eq. (7.90) is just the ***Gauss law***

$$\nabla \cdot \vec{E} = 0. \tag{7.91}$$

[323] In coordinates the Hodge dual of F is $*F_{\mu\nu} = \frac{1}{2} g^{\alpha\rho} g^{\beta\sigma} \varepsilon_{\rho\sigma\mu\nu} F_{\alpha\beta}$, where ε is the Levi-Civita tensor.

To address the undetermination of A_μ we pick a ***gauge condition***, namely the ***Lorenz gauge***[324]

$$G = 0; \qquad G := \partial_\nu A^\nu. \tag{7.92}$$

In terms of this gauge condition, we also introduce the notation

$$R^L_\mu := R_\mu + \partial_\mu G = \Box A_\mu. \tag{7.93}$$

Without imposing any of (7.87), (7.90), or (7.92), the objects R_μ, R^L_μ, C, and G are related by

$$\dot{G} = -C + R^L_0; \tag{7.94}$$

$$\Box G = \partial^\mu R^L_\mu; \tag{7.95}$$

$$\dot{C} = \partial_i R_i = \partial_i (R^L_i - \partial_i G) = \nabla \cdot \vec{R}^L - \Delta G, \tag{7.96}$$

where (7.95) and (7.96) follow from the Bianchi identity (7.89): applying ∂^μ to (7.93) and using (7.89) gives (7.95), whereas (7.96) is (7.89) itself, combined with (7.90) and (7.93).

The point of all this is that instead of directly solving the awkward (i.e. underdetermined as well as overdetermined) Maxwell equations (7.87), one can first solve the wave equation

$$R^L_\mu = 0; \qquad \Leftrightarrow \qquad \Box A_\mu = 0, \tag{7.97}$$

which is of standard hyperbolic type: its solutions for given initial data are even known explicitly. There are two different ways to solve the equations (7.87) via (7.97), which both come down to the simple fact that the conjunction of (7.97) and (7.92) implies (7.87). But they differ in the distribution of labour between (7.97) and (7.92), as follows:

- ***Covariant approach***. Here we solve (7.97) *for each* $\mu = 0,1,2,3$ subject to initial data $A_\mu(\vec{x})$ and $\dot{A}_\mu(\vec{x})$ at $t = 0$ *that respect both the constraint and the gauge condition*:

  $$C(0,\vec{x}) = \Delta A_0(\vec{x}) - \partial_i \dot{A}_i(\vec{x}) = 0; \tag{7.98}$$

  $$G(0,\vec{x}) \equiv \partial_i A_i(\vec{x}) - \dot{A}_0(\vec{x}) = 0. \tag{7.99}$$

  To show that this can indeed be done, first take $A_0(\vec{x}) = \dot{A}_0(\vec{x}) = 0$ (which, incidentally, solves (7.97) by $A_0(x) = 0$), so that (7.98) and (7.99) become $\partial_i \dot{A}_i = 0$ and $\partial_i A_i = 0$, respectively. For example, take $\dot{A}_i(\vec{x}) = 0$ but $A_i(\vec{x}) \neq 0$ arbitrary, and solve the elliptic equation $\Delta\lambda = -\partial_i A_i$ for λ. Replacing A_i by $A_i + \partial_i\lambda$ then satisfies (7.99). Eqs. (7.97), (7.98), and (7.94) imply $\dot{G}(t = 0,\vec{x}) = 0$. Eqs. (7.95) and (7.97) then imply

  $$\Box G = 0. \tag{7.100}$$

  With the initial conditions $G(t = 0,\vec{x}) = 0$, this implies $G(x) = 0$ for all $x \in \mathbb{R}^4$ by the theory of the wave equation. This ***propagation of the gauge*** shows that (7.97) and (7.98) - (7.99) yield (7.87). The analogous ***propagation of the constraint*** $C(t) = 0$ is just a consistency check in this covariant approach: it follows from (7.94), since $G = R^L_0 = 0$, or from (7.96), which implies $\dot{C}(t) = 0$ and, given (7.98), yields $C(t) = 0$ at all t.

- ***Non-covariant approach.*** We solve (7.97) *for each* $\mu = 1,2,3$ only, as well as (7.98) at $t = 0$, but we now have to solve (7.92) for all t. By (7.96) this still gives $\dot{C}(t) = 0$ and hence $C(t) = 0$, so that (7.94) yields $R^L_0 = 0$. We then have (7.97) for $\mu = 0,1,2,3$, and hence, given (7.92), once again have solved (7.87). Note that the propagation of the constraint is independent of the gauge: if $R_i = 0$ whichever way, $C(t) = 0$ follows from the first equality in (7.96) with $C(0) = 0$, and hence from the Bianchi identity.

[324] This gauge should be named after Ludvig Lorenz (1829–1891), rather than H.A. Lorentz (Kragh, 2016).

7.5 General relativity: diffeomorphism invariance and constraints

To start, Einstein's equations (7.1) have two key features analogous to Maxwell's equations:

- They are *underdetermined*: if g solves (7.1), then so does $\psi^* g$, for any $\psi \in \mathrm{Diff}(M)$.
- They are *overdetermined* in that the initial values are constrained (i.e. cannot be arbitrary).

As in the simpler case of electrodynamics, both properties have the same origin, namely the Bianchi identity, here (7.56), i.e. $\nabla^\mu G_{\mu\nu} = 0$. This identity follows from the diffeomorphism invariance of the action (7.2), and it shows that $G_{0\nu}$ contains at most first-order derivatives in time (since $\partial_t G_{0\nu}$ equals some expression containing at most second-order derivatives in time).

The first point also follows from Einstein equations (7.1) themselves, which read

$$G_g = 8\pi T(g,\varphi), \tag{7.101}$$

where G is the Einstein tensor (4.111). From (2.84) with $\psi \rightsquigarrow \psi^{-1}$, (3.54), (4.10) and (4.34) we obtain $R_{\psi^* g} = \psi^* R_g$ (where we explicitly denote the dependence of the Riemann tensor R on the metric g), and similarly for the Ricci tensor, the Ricci scalar, and the Einstein tensor, i.e. $G_{\psi^* g} = \psi^* G_g$. Similarly, the energy-momentum tensor $T(g,\varphi)$ should be constructed such that

$$T(\psi^* g, \psi^* \varphi) = \psi^* T(g,\varphi), \tag{7.102}$$

and hence Einstein's equation (7.101) for g implies

$$G_{\psi^* g} - 8\pi T(\psi^* g, \psi^* \varphi) = \psi^*(G_g - 8\pi T(g,\varphi)) = \psi^* 0 = 0. \tag{7.103}$$

In what follows we just discuss the vacuum case ($T = 0$), since the general case is similar.[325]

From (4.14), (4.15), and (4.109) we easily obtain, in any coordinate system,

$$R_{\mu\nu} = -\tfrac{1}{2} g^{\rho\sigma} g_{\mu\nu,\rho\sigma} - \tfrac{1}{2} g^{\rho\sigma}(g_{\rho\sigma,\mu\nu} - g_{\sigma\nu,\mu\rho} - g_{\mu\rho,\sigma\nu}) + F(g,\partial g), \tag{7.104}$$

where $F(g,\partial g)$ contains only first derivatives of the metric.[326] Anticipating a detailed discussion, we now point out that the ten (vacuum) Einstein equations $G_{\mu\nu} = 0$ come in two groups:

- The six ***dynamical equations*** $G_{ij} = 0$, where $i,j = 1,2,3$ as usual, in which *second*-order time derivatives of the components of the metric occur;
- The four ***constraints*** $C_\mu := G_{\mu 0} = 0$, $\mu = 0,1,2,3$, in which only *first*-order time derivatives of $g_{\mu\nu}$ occur, so that these give relations between initial values for $G_{ij} = 0$.

As in (7.87), the first term in (7.104), which is essentially $\Box g_{\mu\nu}$, has a good PDE theory (as we will see, it makes the spatial components of g satisfy a hyperbolic evolution equation), but the other three terms, which are analogous to the second term in (7.87), ruin this and hence should be removed by a clever choice of coordinates that makes them disappear. The simplest way to do this (introduced by Choquet-Bruhat) is to use the ***wave gauge***,[327] which *given a metric* $g_{\mu\nu}$ is

$$W^\mu = 0; \qquad W^\mu := \Box_g x^\mu, \tag{7.105}$$

where the ***covariant D'Alembertian*** $\Box_g$ is defined, on any tensor, by

$$\Box_g = g^{\rho\sigma}\nabla_\rho \nabla_\sigma. \tag{7.106}$$

[325] The discussion revolves around second derivatives of $g_{\mu\nu}$ in the Einstein equation, which are absent in $T_{\mu\nu}$.
[326] We will later see that in the relevant PDE theory only the highest derivatives of the unknown functions count.
[327] Coordinates satisfying (7.105) are called ***harmonic*** or ***wave coordinates***. See Choquet-Bruhat (2009), §VI.7.

In the wave gauge as defined by condition (7.105), the coordinate functions x^μ are *scalars*,[328] which, once again *given a metric* $g_{\mu\nu}(y)$, are found (locally) as functions $x^\mu(y)$ of some given coordinates (y^μ) by solving $\Box_g x^\mu = 0$ subject to initial conditions on a spacelike hypersurface Σ: if $(\tilde{x}^i)$ are given coordinates on Σ, and N is a fd normal vector field on Σ, we might impose conditions like $x^i_{|\Sigma} = \tilde{x}^i$, $x^0_{|\Sigma} = 0$, $Nx^i = 0$, and $Nx^0 = 1$. *In the new* x^μ*-coordinates*, we have

$$W^\mu = g^{\rho\sigma}\nabla_\rho\partial_\sigma x^\mu = g^{\rho\sigma}(\partial_\rho\partial_\sigma - \Gamma^\nu_{\rho\sigma}\partial_\nu)x^\mu = g^{\rho\sigma}(\partial_\rho\delta^\mu_\sigma - \Gamma^\nu_{\rho\sigma}\delta^\mu_\nu) = -g^{\rho\sigma}\Gamma^\mu_{\rho\sigma}, \quad (7.107)$$

where $g_{\mu\nu}(x)$, and hence $g^{\mu\nu}(x)$ and $\Gamma^\mu_{\rho\sigma}(x)$, are obtained from $g_{\mu\nu}(y)$ using the traditional change of coordinates formula (2.77). Hence (7.105) is a second-order PDE for the x^μ.

Given coordinates (x^μ), on the other hand, the wave gauge (7.105) is seen as a condition on the metric $g_{\mu\nu}(x)$, which because of (7.107) must satisfy any of the equivalent conditions

$$g^{\rho\sigma}\Gamma^\mu_{\rho\sigma} = 0 \quad \Leftrightarrow \quad g^{\rho\sigma}(2g_{\rho\mu,\sigma} - g_{\rho\sigma,\mu}) = 0, \quad \Leftrightarrow \quad \partial_\nu(\sqrt{-\det(g)}g^{\mu\nu}) = 0, \quad (7.108)$$

cf. (7.16), e.g. with corresponding initial conditions $g_{00|S} = -1$ and $g_{0i|S} = 0$. Using (4.15) gives

$$g_{\mu\rho}\partial_\nu W^\rho + g_{\nu\rho}\partial_\mu W^\rho = g^{\rho\sigma}(g_{\rho\sigma,\mu\nu} - g_{\sigma\nu,\mu\rho} - g_{\mu\rho,\sigma\nu}) + H(g,\partial g), \quad (7.109)$$

where $H(g,\partial g)$ has a similar meaning as $F(g,\partial g)$. Therefore, the ***wave-gauged Ricci tensor***

$$R^W_{\mu\nu} \equiv R_{\mu\nu} + \tfrac{1}{2}(g_{\mu\rho}\partial_\nu W^\rho + g_{\nu\rho}\partial_\mu W^\rho), \quad (7.110)$$

cf. (7.97), takes a desirable quasi-linear hyperbolic form, starting with the D'Alembertian:

$$R^W_{\mu\nu} = -\tfrac{1}{2}g^{\rho\sigma}g_{\mu\nu,\rho\sigma} + I(g,\partial g), \quad (7.111)$$

where again I contains only the metric and its first derivatives (though not necessarily linearly).

From (7.110) we also define the ***reduced Einstein tensor***

$$G^W_{\mu\nu} = R^W_{\mu\nu} - \tfrac{1}{2}g_{\mu\nu}R^W = G_{\mu\nu} + \tfrac{1}{2}(g_{\mu\rho}\partial_\nu W^\rho + g_{\nu\rho}\partial_\mu W^\rho - g_{\mu\nu}\partial_\rho W^\rho). \quad (7.112)$$

We then have the following six enlightening analogies between GR and electromagnetism:

$$g_{\mu\nu} \leftrightarrow A_\mu; \qquad W^\mu \leftrightarrow G \qquad C_\mu \leftrightarrow C; \quad (7.113)$$

$$R_{\mu\nu} = 0 \leftrightarrow R_\mu = 0; \qquad R^W_{\mu\nu} = 0 \leftrightarrow R^L_\mu = 0, \qquad \nabla^\mu G_{\mu\nu} = 0 \leftrightarrow \partial_\mu R^\mu = 0. \quad (7.114)$$

Similarly to electromagnetism, there is no good theory for the (vacuum) Einstein equations $R_{\mu\nu} = 0$ we want to solve, whereas there is ample theory for the gauged Einstein equations $R^W_{\mu\nu} = 0$ (though not as explicit and simple as for the wave equation $\Box A_\mu = 0$). In order to follow a similar strategy, we should also find analogues of (7.94) - (7.96). First, we have

$$G^W_{\mu\nu} = R^W_{\mu\nu} - \tfrac{1}{2}g_{\mu\nu}R^W = G_{\mu\nu} + \tfrac{1}{2}(g_{\mu\rho}\partial_\nu W^\rho + g_{\nu\rho}\partial_\mu W^\rho - g_{\mu\nu}\partial_\rho W^\rho), \quad (7.115)$$

so that, taking $\nu = 0$, eq. (7.94) is replaced by four equations ($\mu = 0,1,2,3$)

$$\tfrac{1}{2}(g_{\mu i}\dot{W}^i + g_{00}\partial_\mu W^0 + g_{0i}\partial_\mu W^i - g_{\mu 0}\partial_i W^i) = -C_\mu + G^W_{\mu 0}. \quad (7.116)$$

[328] As opposed to components of a 4-vector. Choquet-Bruhat even writes $x^{(\mu)}$ as a warning.

The analogue of (7.95) follows by applying the Bianchi identity (7.56) to (7.115). Using the fact that the W^ρ are scalars and the Levi-Civita connection ∇ is metric and torsion-free, we compute

$$\nabla^\mu(g_{\mu\rho}\partial_\nu W^\rho - g_{\mu\nu}\partial_\rho W^\rho) = \nabla_\rho\partial_\nu W^\rho - \nabla_\nu\partial_\rho W^\rho = (\partial_\rho\partial_\nu - \partial_\nu\partial_\rho + \Gamma^\sigma_{\rho\nu} - \Gamma^\sigma_{\nu\rho})W^\rho = 0.$$

Hence eqs. (7.56) and (7.115) give

$$\Box_g W^\mu = 2g^{\mu\rho}\nabla^\nu G^W_{\rho\nu}. \tag{7.117}$$

Finally, the counterpart of (7.96) in GR again follows from the Bianchi identity (7.56), viz.

$$\partial^0 C_0 = -\partial^j C_j + (g^{\rho\sigma}\Gamma^\mu_{\rho\sigma} + g^{0\rho}\Gamma^\mu_{\rho 0})C_\mu + g^{j\rho}(\Gamma^0_{\rho 0}C_j + \Gamma^k_{\rho 0}G_{jk}); \tag{7.118}$$

$$\partial^0 C_i = g^{\rho\sigma}(\Gamma^0_{\rho\sigma}C_i + \Gamma^0_{\rho i}C_\sigma) + g^{\rho 0}\Gamma^j_{\rho i}C_j - \partial^j G_{ij} + g^{\rho\sigma}\Gamma^j_{\rho\sigma}G_{ij} + g^{\rho k}\Gamma^j_{\rho i}G_{jk}, \tag{7.119}$$

where we may also write G_{ij} in terms of G^W_{ij} and W^μ via (7.115), e.g. for (7.118) this gives

$$(\partial^\mu + W^\mu)C_\mu = g^{\sigma\nu}\Gamma^\mu_{0\sigma}(G^W_{\mu\nu} - \tfrac{1}{2}(g_{\mu\rho}\partial_\nu W^\rho + g_{\nu\rho}\partial_\mu W^\rho - g_{\mu\nu}\partial_\rho W^\rho)). \tag{7.120}$$

Knowing all this, we can solve the vacuum Einstein equations $G_{\mu\nu} = 0$ in two alternative ways:

- ***Covariant approach***. We solve the covariant (*space-time*) reduced Einstein equations

$$R^W_{\mu\nu} = 0 \tag{7.121}$$

for all values $\mu,\nu = 0,1,2,3$. This can indeed be done, because (7.121) with (7.110) is a hyperbolic quasi-linear PDE system for which good existence, uniqueness, and stability results exist; see §7.6. Since (7.121) gives $g^{\mu\nu}R^W_{\mu\nu} = 0$, it also implies $G^W_{\mu\nu} = 0$. Furthermore, we impose both the constraints and the gauge conditions at $t = 0$, i.e.,

$$C_\mu(t = 0,\vec{x}) = 0; \tag{7.122}$$

$$W^\mu(t = 0,\vec{x}) = 0 \tag{7.123}$$

Then (7.116) also gives

$$\dot{W}^\mu(t = 0) = 0, \tag{7.124}$$

upon which (7.117) gives $W^\mu(t) = 0$ at all t, since this is the unique solution with initial data (7.123) and (7.124). This is the ***propagation of the gauge***. The full Einstein equations $R_{\mu\nu} = 0$ are then satisfied because of (7.110), and finally–though unnecessary in this approach–propagation of the constraints may be verified from (7.118) - (7.119).

- ***Non-covariant approach***. We solve only the *spatial* part of the reduced Einstein equations

$$G^W_{ij} = 0 \qquad\qquad (i,j = 1,2,3), \tag{7.125}$$

whilst imposing the initial value constraints (7.122) *at* $t = 0$, and the *full* gauge condition $W^\mu(x) = 0$ *for all* $x = (t,\vec{x})$. Since this gives $G^W_{\mu\nu} = G_{\mu\nu}$, the Bianchi identities (7.118) - (7.119) with $G_{ij} = 0$ simply become

$$\partial^0 C_0 = (-\partial^j + g^{j\rho}\Gamma^0_{\rho 0})C_j + (g^{\rho\sigma}\Gamma^\mu_{\rho\sigma} + g^{0\rho}\Gamma^\mu_{\rho 0})C_\mu; \tag{7.126}$$

$$\partial^0 C_i = g^{\rho\sigma}(\Gamma^0_{\rho\sigma}C_i + \Gamma^0_{\rho i}C_\sigma) + g^{\rho 0}\Gamma^j_{\rho i}C_j. \tag{7.127}$$

Thus the constraints satisfy coupled homogeneneous first-order hyperbolic PDEs, whose solution with given initial condition (7.122) is zero. This ***propagation of the constraints*** via the Bianchi identity gives the remaining Einstein equations $G_{\mu 0} = 0$, and since we already had $G_{ij} = 0$ for $i,j = 1,2,3$ from $G^W_{ij} = 0$ and $W^\mu = 0$, we seem ready!

There is a complication, however, in that the reduced Einstein equations (7.125) are neither *a priori* in suitable (i.e. hyperbolic) form, nor do they follow from $R^W_{ij} = 0$ (which *are* in suitable form), because the Einstein tensor $G_{\mu\nu}$ also involves the Ricci scalar R, which cannot be computed from R^W_{ij} alone. This can be resolved by passing to the (inverse) ***densitized metric***

$$\mathfrak{g}^{\mu\nu} = g^{\mu\nu}\sqrt{|\det(g)|}, \tag{7.128}$$

in terms of which the gauge condition and the gauged Einstein equations read

$$\partial_\nu \mathfrak{g}^{\mu\nu} = 0; \tag{7.129}$$
$$G^{\mu\nu}_W = 0, \tag{7.130}$$

and moreover the gauged Einstein tensor (7.115) turns out to take the desired hyperbolic form

$$|\det(g)|G^{\mu\nu}_W = \tfrac{1}{2}\mathfrak{g}^{\rho\sigma}\partial_\rho\partial_\sigma\mathfrak{g}^{\mu\nu} + O(\mathfrak{g},\partial\mathfrak{g}). \tag{7.131}$$

We will not follow this path, but will set up the non-covariant approach in a more geometric way in §7.7. This will still follow the general idea of solving $R^W_{ij} = 0$ or $G^W_{ij} = 0$ with *initial value constraints* (7.122), and a *full* (but non-covariant) gauge condition like $W^\mu(x) = 0$.

Finally, since we have already treated electromagnetism as a warm-up for GR, it is interesting to combine the two in the light of gauge fixing and constraints. Hence we briefly study the coupled Einstein–Maxwell equations (7.62) with (7.84) and (7.79). Since $T = 0$, these become

$$R_{\mu\nu} = 2(g^{\rho\sigma}F_{\mu\rho}F_{\nu\sigma} - \tfrac{1}{4}g_{\mu\nu}F^2); \tag{7.132}$$
$$R_\mu := \nabla^\nu F_{\nu\mu} = 0. \tag{7.133}$$

Everything in §7.4 goes through, provided we replace ordinary derivatives by covariant ones, as in (7.133). For example, in the derivation of the Bianchi identity eq. (7.88) becomes

$$0 = \int_V d^4x\sqrt{-g}\,\nabla_\nu F^{\nu\mu}\partial_\mu\lambda = -\int_V d^4x\sqrt{-g}\,\lambda\nabla_\mu\nabla_\nu F^{\nu\mu}, \tag{7.134}$$

where ∇_μ instead of ∂_μ arises because of (7.17). Hence the Bianchi identity (7.89) becomes

$$\nabla_\mu R^\mu = 0, \tag{7.135}$$

which unlike (7.89) is far from trivial. A simple computation using (4.13) and (4.109) yields

$$R_\mu = \Box_g A_\mu - R_{\mu\nu}A^\nu - \partial_\mu G, \tag{7.136}$$

where, in the spirit of what was just said, the covariant Lorenz gauge is now given by

$$G = \nabla_\nu A^\nu. \tag{7.137}$$

Putting $R^L_\mu = R_\mu + \partial_\mu G$ as before, we have $R^L_\mu = \Box_g A_\mu - R_{\mu\nu}A^\nu$. In order to solve (7.132) - (7.133), then, we must solve the gauge conditions $G = 0$ and $W^\mu = 0$ and the hyperbolic system

$$R^W_{\mu\nu} = 2(g^{\rho\sigma}F_{\mu\rho}F_{\nu\sigma} - \tfrac{1}{4}g_{\mu\nu}F^2); \tag{7.138}$$
$$\Box_g A_\mu = R_{\mu\nu}A^\nu. \tag{7.139}$$

Spelling out the "covariant" and "non-covariant" approaches is now just a tedious exercise.

7.6 Existence, uniqueness, and maximality of solutions

In this section we give a geometric formulation of the Cauchy problem for the Einstein equations, including its (abstract) solution in the form of Theorem 7.10, obtained in 1969 by Choquet-Bruhat and Geroch (following two decades of progress mainly due to Choquet-Bruhat).[329]

So far, the procedure in §7.5 leads to solutions that are *local in space and local in time*:

- Locality in *space* follows from the use of specific coordinates, i.e. those satisfying (7.105).
- Locality in *time* is all that the existence (and uniqueness and stability) theorems for quasi-linear second-order hyperbolic PDEs of the kind (7.121) provide.

We now indicate how this can be improved. First, local existence in *space* turns into global existence in space by globalizing the gauge, as follows. A well-known concept in Riemannian geometry is that of a ***harmonic map*** $h : M \to \hat{M}$ between Riemannian manifolds (M, g) and $(\hat{M}, \hat{g})$. These maps can be described abstractly, but is is easier to use local coordinates (x^μ) on M, and likewise $(\hat{x}^i)$ on $\hat{M}$. Any map $h : M \to \hat{M}$ has an associated *energy* functional, defined by

$$E(h) := \int_M d^3x \sqrt{g(x)} e_x(h); \qquad e_x(h) := \tfrac{1}{2} g^{\mu\nu}(x) \hat{g}_{ij}(h(x)) \frac{\partial h^i(x)}{\partial x^\mu} \frac{\partial h^j(x)}{\partial x^\nu}, \tag{7.140}$$

where h^i are the components of h relative to the coordinates $(\hat{x}^i)$. This expression turns out to be independent of the coordinates.[330] For example, if $M = [a, b]$ with flat metric, then $E(f)$ is the energy (3.23) of a curve in N. Another example is $N = \mathbb{R}$ with flat metric, in which case

$$E(h) = \int_M \nabla h \cdot \nabla h \tag{7.141}$$

is the ***Dirichlet integral*** of h, which plays a fundamental role in the theory of the Laplace equation $\Delta h = 0$ on M. It can be shown that h extremizes $E(h)$ iff it solves the equation

$$g^{\mu\nu} \left(\frac{\partial^2 h^i(x)}{\partial x^\mu \partial x^\nu} - \Gamma^\rho_{\mu\nu}(x) \frac{\partial h^i(x)}{\partial x^\rho} + \hat{\Gamma}^i_{jk}(h(x)) \frac{\partial h^j(x)}{\partial x^\mu} \frac{\partial h^k(x)}{\partial x^\nu} \right) = 0, \tag{7.142}$$

where and $\Gamma^\rho_{\mu\nu}$ and $\hat{\Gamma}^i_{jk}$ are the Christoffel symbols for g and $\hat{g}$, respectively. Thus h is called *harmonic* if it solves (7.142). Exactly the same constructions work in Lorentzian geometry, in which case a solution of (7.142) is called a ***wave map***. In that case, standard hyperbolic PDE theory yields existence and uniqueness of solutions $h_{|\Sigma}$ and $\dot{h}_{|\Sigma}$ subject to initial conditions on a Cauchy surface Σ in M, which we (evidently) assume to be globally hyperbolic.

In order to provide the right version of the wave gauge enabling global solutions in space, we pick some fiducial Riemannian metric γ on our Σ and introduce the Lorentzian manifold

$$\hat{M} := \mathbb{R} \times \Sigma \qquad\qquad \hat{g} := -dt^2 + \gamma. \tag{7.143}$$

Definition 7.3 *We say that a (Lorentzian) metric g on $M = \mathbb{R} \times \Sigma$ satisfies the $\hat{g}$-wave gauge iff the identity map* $\mathrm{id} : M \to \hat{M}$ *is a wave map with respect to g and $\hat{g}$.*

[329] Introductions to the Cauchy problem in GR, from different perspectives, include Choquet-Bruhat & York (1980), Friedrich & Rendall (2000), Klainerman & Nicolò (2003), Rendall (2005, 2008), Christodoulou (2008), Dafermos (2009), Choquet-Bruhat (2009), Ringström (2009), Chruściel (2010), and Aretakis & Rodnianski (2015).

[330] See e.g. Jost (2002), §8.1.

It follows from the coordinate-independence of (7.142) that this condition is coordinate-independent. One can also see this explicitly by noting that g satisfies the $\hat{g}$-wave gauge iff

$$\hat{W}^\mu = 0 \tag{7.144}$$

for each $\mu = 0,1,2,3$, where, cf. (7.105) and (7.107),

$$\hat{W}^\mu = g^{\rho\nu}(\hat{\Gamma}^\mu_{\rho\nu} - \Gamma^\mu_{\rho\nu}). \tag{7.145}$$

Since the difference between two connections is a tensor (see §7.2), the index μ is now a true vector index in that $\hat{W}^\mu$ are the components of a vector. Thus the coordinate-dependence of the original wave gauge has been traded for $\hat{g}$-dependence. We now follow the same steps as for the wave gauge, replacing W by $\hat{W}$ from (7.109) till the end of §7.5, with the same conclusions: the reduced Einstein equations are quasi-linear and hyperbolic, the gauge and the constraints propagate, etc., with the difference that none of the arguments now depend on the choice of local coordinates and hence local (coordinate) solutions can be patched together so as to become globally defined in space, that is, on Σ. In particular, we may treat the μ in (7.145) as a vector index and write down neat covariant formulae. This gives, for example,

$$R^{\hat{W}}_{\mu\nu} := R_{\mu\nu} + \tfrac{1}{2}(\nabla_\mu \hat{W}_\nu + \nabla_\nu \hat{W}_\mu) = -\tfrac{1}{2} g^{\rho\sigma} g_{\mu\nu,\rho\sigma} + \hat{I}(g, \partial g); \tag{7.146}$$

$$\Box_g \hat{W}_\mu + R^\nu_\mu \hat{W}_\nu = \nabla^\nu G^{\hat{W}}_{\mu\nu}, \tag{7.147}$$

cf. (7.110) - (7.112) and (7.117), which still have a desirable hyperbolic form. *Mutatis mutandis*, both the covariant and the non-covariant approaches of the previous section may then proceed.

In order to (at least partially) overcome the problems with locality in *time* mentioned above we explain a specific way of posing the initial data that–within PDE theory–seems unique for GR. This construction not only brings the initial data in geometric form (as opposed to giving $(g_{\mu\nu}(t=0), \dot{g}_{\mu\nu}(t=0))$ as might expected for hyperbolic PDEs) but also solves the closely related problem that naively a solution (M,g) to the Einstein equations would be based on initial data given on some Cauchy surface $\Sigma \subset M$ where M is given; but the problem in GR is that the manifold M is typically constructed along with the metric g, as opposed to be given *in advance*.

To find the correct geometric way of posing the Cauchy problem for GR, we first *assume* we have a globally hyperbolic space-time (M,g) solving the Einstein equations (in vacuum or with matter), *assume* we have a spacelike Cauchy surface $\Sigma \subset M$, seen as a triple $(M,\Sigma,\imath)$, where $\imath : \Sigma \hookrightarrow M$ injects some given 3-manifold Σ into M (as an embedded submanifold, cf. Definition 4.13), and figure out which initial data the triple $(M,g,\imath)$ puts on Σ. These initial data will then be taken by themselves, after which the ambient space-time (M,g) can be forgotten.

As already mentioned, instead of $g_{\mu\nu}$ and $\dot{g}_{\mu\nu}$ at Σ, one prefers geometric data, namely:

- The induced Riemannian 3-metric $\tilde{g} := \imath^* g$, cf. (4.124);
- The extrinsic curvature $\tilde{k}$ of the embedding $\imath : \Sigma \hookrightarrow M$, see (4.144).

In the next section (§7.7) we will show that the Einstein equations impose constraints on these quantities (see also §7.5 for motivation and context), which in the vacuum case are

$$\tilde{R} - \mathrm{Tr}\,(\tilde{k}^2) + \mathrm{Tr}\,(\tilde{k})^2 = 0; \tag{7.148}$$

$$\tilde{\nabla}_j \tilde{k}^j_i - \tilde{\nabla}_i \mathrm{Tr}\,(\tilde{k}) = 0. \tag{7.149}$$

Here $\tilde{R}$ is the Ricci scalar on Σ for the Riemannian metric $\tilde{g}$ and likewise $\tilde{\nabla}$ is the Levi-Civita connection on Σ determined by $\tilde{g}$. Thus the initial data for the Einstein equations are triples

$$(\Sigma, \tilde{g}_{ij}, \tilde{k}_{ij}) \equiv (\Sigma, \tilde{g}, \tilde{k}), \tag{7.150}$$

subject to the vacuum constraints (7.148) - (7.149), or their matter analogues (8.65) - (8.67).

Definition 7.4 *Given initial data $(\Sigma, \tilde{g}, \tilde{k})$ satisfying the constraints (7.148) - (7.149), any triple $(M, g, \imath)$ that solves the Einstein equations and induces these initial data in such a way that $\imath(\Sigma)$ is a Cauchy surface in M, so that in particular (M, g) is globally hyperbolic, is called a* **Cauchy development** *or* **globally hyperbolic development** *of the data $(\Sigma, \tilde{g}, \tilde{k})$.*

The theorems below may then be summarized as follows:

Theorem 7.5 *Let $(\Sigma, \tilde{g})$ be a 3d Riemann manifold equipped with a second symmetric tensor*

$$\tilde{k} \in \mathfrak{X}^{(2,0)}(\Sigma)$$

such that $(\Sigma, \tilde{g}, \tilde{k})$ satisfies the constraints (7.148) - (7.149). Then there exists a maximal globally hyperbolic space-time (M, g) and an isometric embedding $\imath : \Sigma \hookrightarrow M$ for which the extrinsic curvature is the given $\tilde{k}$, and such a space-time is unique up to isometry.

We will explain what 'maximal' means here. It is interesting to compare this with Theorem 4.18, i.e. the fundamental theorem for hypersurfaces, which for this purpose we rephrase as follows:

Theorem 7.6 *Let $(\Sigma, \tilde{g})$ be a connected and simply connected Riemann manifold equipped with a second symmetric tensor*

$$\tilde{k} \in \mathfrak{X}^{(2,0)}(\Sigma)$$

such that $(\Sigma, \tilde{g}, \tilde{k})$ satisfies the Gauss–Codazzi equations

$$\tilde{R}_{ijkl} + \tilde{k}_{il}\tilde{k}_{jk} - \tilde{k}_{ik}\tilde{k}_{jl} = 0; \tag{7.151}$$

$$\tilde{\nabla}_i \tilde{k}_{jk} - \tilde{\nabla}_j \tilde{k}_{ik} = 0. \tag{7.152}$$

If $m = \dim(\Sigma) \geq 2$, there exists an isometric embedding $\imath : \Sigma \to \mathbb{R}^{m+1}$ for which the extrinsic curvature is the given tensor $\tilde{k}$, and such an embedding is unique up to Euclidean motions (i.e. up to isometries, which are combinations of translations and rotations).

The constraints (7.151) - (7.152) are stronger than (7.148) - (7.149); up to a relative sign, which accounts for the difference between the Riemannian and the Lorentzian cases, see (4.148), eq. (7.148) follows from (7.151) by contracting it with $\tilde{g}^{ik}\tilde{g}^{jl}$, whilst (7.149) follows from (4.155) by contracting with $\tilde{g}^{ik}$. The reason is that Theorem (4.18) asks for a stronger result, namely embedding into Euclidean space, where $R_{\rho\sigma\mu\nu} = 0$, whereas Theorem 7.5 merely asks for embedding in a Lorentzian manifold where $R_{\mu\nu} = 0$. Otherwise, the spirit of the two theorems is similar, in that the Gauss–Codazzi equations and the constraints in GR, whose geometric form (7.148) - (7.149) will actually be derived from the Gauss–Codazzi equations, both arise as consistency conditions for the existence of a certain embedding of the initial data set $(\Sigma, \tilde{g}, \tilde{k})$:

- into Euclidean space in the nineteenth-century fundamental theorem for hypersurfaces;
- into a space-time solving the Einstein equations in the twentieth-century Theorem 7.5.

We will now dissect Theorem 7.5, in particular making precise (in steps) what it means for the space-time (M,g) to be maximal (this will be done in the crowning Theorem 7.10).[331]

Theorem 7.7 *For any smooth initial data set $(\Sigma,\tilde{g},\tilde{k})$ satisfying the constraints (7.148) - (7.149) there is an open interval $0\in I\subset\mathbb{R}$ and a Lorentzian metric g on $M=I\times\Sigma$ such that $(M,g,\imath)$, where $\imath:\Sigma\hookrightarrow M$ given by $\imath(\vec{x})=(0,\vec{x})$, is a Cauchy development of $(\Sigma,\tilde{g},\tilde{k})$. Moreover, (M,g) is automatically a globally hyperbolic space-time, with Cauchy surface $\imath(\Sigma)$.*

In §7.5 we have only sketched the part of the existence proof that reduces the Einstein equations to a simpler problem involving quasilinear hyperbolic PDEs, whose theory we briefly review in Appendix B; the entire proof can be found in the literature.[332] We now turn to uniqueness.

Theorem 7.8 **(Geometric uniqueness of solutions of Einstein's equations)** *Any two Cauchy developments $(M_1,g_1,\imath_1)$ and $(M_2,g_2,\imath_2)$ of the same (smooth) initial data are locally isometric, in that $\imath_1(\Sigma)$ and $\imath_2(\Sigma)$ have open neighbourhoods U_1 and U_2 in M_1 and M_2, respectively, such that (U_1,g_1) and (U_2,g_2) are isometric through a diffeomorphism $\psi:U_1\to U_2$ satisfying*

$$\psi^*g_2=g_1;\qquad\qquad \psi\circ\imath_1=\imath_2. \tag{7.153}$$

The proof is very involved,[333] but the idea is the argument for underdeterminacy explained at the beginning of §7.5, where we require ψ to preserve the initial data (this is Hilbert's version of Einstein's Hole Argument, cf. §1.5). Technically, construct wave maps $h_i:\hat{M}\to M_i$ $(i=1,2)$, suitably shrunk to as to become diffeomorphisms, and define $g_i'=h_i^*g_i$ on $\hat{M}$. This brings both g_1 and g_2 into the $\hat{g}$-wave gauge. These new metrics solve the same equations, namely the reduced Einstein equations and the $\hat{g}$-wave gauge condition, with the same initial conditions. Hence they must coincide by local uniqueness result from hyperbolic PDEs. From $g_1'=g_2'$ we then obtain

$$g_2=(h_1^{-1}\circ h_2)^*g_1=\psi^*g_1. \tag{7.154}$$

Definition 7.9 *A* **maximal Cauchy development** *or $(M_{\max},g_{\max},\imath_{\max})$ of given (smooth) initial data $(\Sigma,\tilde{g},\tilde{k})$ satisfying the constraints (7.148) - (7.149) is a Cauchy development with the property that for any other Cauchy development $(M,g,\imath)$ of these data there exists an embedding*

$$\psi:M\to M_{\max}$$

that preserves time orientation, metric, and Cauchy surface, i.e., one has

$$\psi^*g_{\max}=g;\qquad\qquad \psi\circ\imath=\imath_{\max}. \tag{7.155}$$

Compare with (7.153). Since a maximal Cauchy development is always globally hyperbolic, it is also called a ***maximal globally hyperbolic development*** or MGHD of the initial data.

The word "maximal" is confusing. It does not imply that $(M_{\max},g_{\max},\imath_{\max})$ is maximal *as a solution to the vacuum Einstein equations with given initial data*, or *as a space-time*. It does not even mean that $(M_{\max},g_{\max})$ cannot have any globally hyperbolic extensions. It *does* mean that:

[331] We only discuss smooth initial data. See footnote 338 for the non-smooth case. As shown in Appendix B, the smooth case is proved from the case with initial data and thence solutions in Sobolev spaces H^s and letting $s\to\infty$.

[332] See e.g. Choquet-Bruhat (2009), chapter VI and Appendix III, and in Ringström (2009), chapter 14.

[333] See e.g. Choquet-Bruhat (2009), Theorem VI.8.4, or Ringström (2009), Theorem 14.3.

If $(M_{\max}, g_{\max}, \imath_{\max})$ can be (properly) isometrically embedded in some space-time (M', g'), then the ensuing copy of Σ in M' (arising from $\Sigma \hookrightarrow M \hookrightarrow M'$) cannot be a Cauchy surface in M'.

In particular, $\Sigma \subset M'$ would have a nonempty Cauchy horizon. This is often taken to indicate an end to determinism, but this seems an overstatement. The correct statement is that the existence of a Cauchy horizon for Σ in the extension M' means that (M', g'), unlike (M, g), is no longer *predictable from initial data on* Σ. It may in principle be predictable from some new Cauchy surface Σ' that is not the image of any Cauchy surface in M under the embedding, although in typical examples (cf. §10.6) the larger space-time (M', g') is in fact not globally hyperbolic.[334] Strong cosmic censorship (in its current formulation, which is different from Penrose's original one) excludes such extensions, and so we will take up this topic in more detail in §§10.4–10.5.

We now come to the main abstract result in the initial-value approach to GR.

Theorem 7.10 (Choquet-Bruhat and Geroch) *Each smooth initial data set $(\Sigma, \tilde{g}, \tilde{k})$ satisfying the constraints has a maximal Cauchy development $(M_{\max}, g_{\max}, \imath_{\max})$, which is unique up to time-orientation-preserving isometries fixing the Cauchy surface $\imath(\Sigma) \subset M_{\max}$, as in (7.153).*

For understanding both the claim and its proof it is useful to rephrase Theorem 7.10 in terms of partially ordered sets (posets). We already saw that Cauchy developments of fixed initial data are far from unique due to diffeomorphism invariance of the Einstein equations. We circumvent this apparent lack of determinism by declaring two solutions equivalent if they can be transformed onto each other by a diffeomorphism respecting $\imath$ as well as time orientation. Thus we say that

$$(M_1, g_1, \imath_1) \cong (M_2, g_2, \imath_2) \tag{7.156}$$

iff there is a time-orientation preserving diffeomorphism $\psi : M_1 \to M_2$ satisfying (7.153). This is an equivalence relation on the set $\mathrm{GHD}(\Sigma, \tilde{g}, \tilde{k})$ of all globally hyperbolic (i.e. Cauchy) developments of the data $(\Sigma, \tilde{g}, \tilde{k})$. We denote the (quotient) set of its equivalence classes by $[\mathrm{GHD}](\Sigma, \tilde{g}, \tilde{k})$. As usual, we write $[M, g, \imath]$ for the equivalence class of $(M, g, \imath)$.

Definition 7.11 *Initially, put*

$$(M_1, g_1, \imath_1) \leq (M_2, g_2, \imath_2) \tag{7.157}$$

iff there is a embedding $\psi : M_1 \to M_2$ *such that (7.153) hold. This fails to be a partial ordering on $\mathrm{GHD}(\Sigma, \tilde{g}, \tilde{k})$ (it fails the antisymmetry axiom), but it does descend to a partial ordering on $[\mathrm{GHD}](\Sigma, \tilde{g}, \tilde{k})$. By abuse of notation, provided (7.157) holds we may therefore write*

$$[M_1, g_1, \imath_1] \leq [M_2, g_2, \imath_2]. \tag{7.158}$$

This makes $([\mathrm{GHD}](\Sigma, \tilde{g}, \tilde{k}) \leq)$ a partially ordered set (poset).

Recall that a ***top element*** $\top \in P$ of a poset $(P, \leq)$ is an element for which $x \leq \top$ for all $x \in P$. A top element need not exist, but it is unique if it exists.[335] Theorem 7.10 then becomes:

Theorem 7.12 *The poset $([\mathrm{GHD}](\Sigma, \tilde{g}, \tilde{k}) \leq)$ has a top element (which is necessarily unique).*

[334] See Doboszewski (2017, 2019, 2020) and Manchak (2011, 2017) for studies of the (in)extendibility of space-times, partly in connection global hyperbolicity.

[335] This is different from a *maximal element* $m \in P$, where for all $x \in P$ one has $m \leq x$ iff $x = m$. Maximal elements are often non-unique if they exist, and even if they are unique they may not be top elements (which are maximal).

Note that Theorem 7.12 *implies* that the maximal Cauchy development $(M_{\max}, g_{\max}, \imath_{\max})$ is unique up to isometry.[336] We first rephrase Theorem 7.8 in terms of the above poset:

Corollary 7.13 *Any two Cauchy developments $(M_1, g_1, \imath_1)$ and $(M_2, g_2, \imath_2)$ of given initial data have a common Cauchy development $(M, g, \imath)$, in that in that we have both orderings*

$$(M, g, \imath) \leq (M_1, g_1, \imath_1); \qquad (M, g, \imath) \leq (M_2, g_2, \imath_2). \tag{7.159}$$

Indeed, take $M = U_1$, with:

- $\psi_1 : M \to M_1$ given by the embedding $i : U_1 \subset M_1$;
- $\psi_2 : M \to M_2$ defined by $\psi_2 = \psi \circ i$, where ψ is the map from Theorem 7.8.

More strongly, we even have:

Lemma 7.14 *Any two Cauchy developments $(M_1, g_1, \imath_1)$ and $(M_2, g_2, \imath_2)$ have a* maximal *common Cauchy development $(M', g', \imath')$, in that any other common Cauchy development satisfies*

$$(M, g, \imath) \leq (M', g', \imath'). \tag{7.160}$$

Indeed, if $\{U_\alpha\}$ is the set of all U_1's appearing in Theorem 7.8, i.e. $U_\alpha \subset M_1$ with given maps $\psi_\alpha : U_\alpha \to M_2$, etc., then one may simply take the union $M' = \cup_\alpha U_\alpha$, with the obvious embedding $M' \subset M_1$, and the map $\psi : M' \to M_2$ given by $\psi(x) = \psi_\alpha(x)$ if $x \in U_\alpha$. Conversely:

Lemma 7.15 *Any two Cauchy developments $(M_1, g_1, \imath_1)$ and $(M_2, g_2, \imath_2)$ have a common* extension *$(M_{12}, g_{12}, \imath_{12})$, in that we have both orderings*

$$(M_1, g_1, \imath_1) \leq (M_{12}, g_{12}, \imath_{12}); \qquad (M_2, g_2, \imath_2) \leq (M_{12}, g_{12}, \imath_{12}). \tag{7.161}$$

Define an equivalence relation $\sim$ on the disjoint union $M_1 \sqcup M_2$ of M_1 and M_2 by $x \sim y$ if:

- either $x = y$;
- or $x \in M' \subset M_1$ and $y = \psi(x)$, where $\psi : M' \to M_2$ has just been defined.

The quotient

$$M_{12} = (M_1 \sqcup M_2) / \sim \tag{7.162}$$

inherits a metric g_{12} from (M_1, g_1) and (M_2, g_2), as follows:

- for $x \in M_1 \backslash M'$ we put $g_{12}([x]) := g_1(x)$;
- for $y \in M_2 \backslash \psi(M')$ take $g_{12}([y]) := g_2(y)$, noting that $[x] = x$ and $[y] = y$ in those cases;
- for $x \in M_1$ and $y = \psi(x)$, so that $[x] = [y]$, we put $g_{12}([x]) := g_1(x)$ $(= g_2(y))$.

[336] Choquet-Bruhat & Geroch (1969) sketched a proof based on Zorn's lemma, which they even had to use twice. The corresponding proof in Ringström (2009), §14, is wrong, but is corrected in Ringström (2013), §23. Instead, we outline a recent constructive proof due to Sbierski (2016), with some improvements by Wong (2013).

The obvious maps $M_1 \hookrightarrow M_{12}$ and $M_2 \hookrightarrow M_{12}$ are isometries for g_{12} by construction.[337] Similarly, we obtain embeddings $\Sigma \hookrightarrow M_{12}$ and $\Sigma \hookrightarrow M_{12}$ from the given ones $\Sigma \hookrightarrow M_1$ and $\Sigma \hookrightarrow M_2$.

The construction of the maximal space-time $M_{\max}$ is an extension of (7.162). One defines

$$M_{\max} = (\sqcup_\lambda M_\lambda) \,/\, \sim, \tag{7.163}$$

where $\{M_\lambda\}$ is the set of all Cauchy developments (of the given initial data), and we identify $x \in M_1$ and $y \in M_2$ (where 1 and 2 are generic values of λ) iff $x \sim y$ as defined after (7.162). Also, the constructions of the metric $g_{\max}$, the embedding $\imath_{\max}$, and the (isometric) embeddings

$$\psi_\lambda : M_\lambda \to M_{\max} \tag{7.164}$$

are entirely similar to the case (7.162) just explained. Maximality is then obvious. □

Theorem 7.10 is stated for smooth initial data, which give rise to smooth 4-metrics. However, the local existence results whose proof we omitted are proved by taking limits of existence results for rougher initial data in Sobolev spaces $H^s(\Sigma)$ as $s \to \infty$ (see Appendix B.3 for notation and details). These lower regularity results are also of interest as such. In particular, we have:[338]

Theorem 7.16 *Let $s > 3/2$. For initial data $(\Sigma, \tilde{g}_{ij}, \tilde{k}_{ij})$ where $\tilde{g}$ is sufficiently close to $\hat{\gamma}$ and*

$$\tilde{g} \in H^{s+1}_{(2,0)}(\Sigma); \qquad \tilde{k} \in H^s_{(2,0)}(\Sigma), \tag{7.165}$$

there is $T > 0$ such that the reduced vacuum Einstein equations (7.121) or their counterparts in a $\hat{g}$-wave gauge, have a unique solution g on $M = [0,T] \times \Sigma$, where

$$g_{\mu\nu} \in C([0,T], H^{s+1}(\Sigma)) \cap C^1([0,T], H^s(\Sigma)); \tag{7.166}$$

$$\partial_\rho g_{\mu\nu} \in C([0,T], H^s(\Sigma)). \tag{7.167}$$

This solution continuously depends on the initial data, in that $\tilde{g}_l \to \tilde{g}$ in $H^{s+1}_{(2,0)}(\Sigma)$ and $\tilde{k}_l \to \tilde{k}$ in $H^s_{(2,0)}(\Sigma)$ imply $g_l \to g$ in $L^\infty([0,T], H^{s+1}(\Sigma))$ as well as $\partial_\rho g_l \to g$ in $L^\infty([0,T], H^s(\Sigma))$.

For $s > m + 3/2$, the Sobolev embedding theorem (B.23) gives $H^s(\Sigma) \subset C^m(\Sigma)$, so that for $s > 3/2$ and hence $m = 0$, eq. (7.165) imply that $\tilde{g} \in C^1(\Sigma)$ and $\tilde{k} \in C(\Sigma)$, upon which eqs. (7.166) - (7.167) then imply $g \in C^1(M)$ and hence $\partial g \in C(M)$. Another refinement is *localization*. For example, Theorem 7.8 gives rise to what is best seen as a *causality* result:

Proposition 7.17 *Let $(\tilde{g}_{ij}, \tilde{k}_{ij})$ and $(\tilde{g}'_{ij}, \tilde{k}'_{ij})$ be (smooth) initial data on Σ that coincide on some submanifold $\Sigma_0 \subset \Sigma$. Then any two Cauchy developments $([0,T] \times \Sigma, g)$ and $([0,T'] \times \Sigma, g')$ of these data are isometric when restricted to $D^+(\Sigma_0) \subset [0,T''] \times \Sigma_0$, where $T'' = \min\{T, T'\}$.*

[337] The main difficulty in the proof is to show that M_{12} is a Hausdorff space; see the references in footnote 336.

[338] Here $\hat{\gamma}$ is a fiducial Riemannian metric on Σ enabling a coordinate-independent definition of Sobolev spaces on Σ. The index $(2,0)$ in $H^s_{(2,0)}(\Sigma)$ refers to the tensor character of $\tilde{g}$ and $\tilde{k}$; one has such Sobolev spaces for any (k,l). Choquet-Bruhat's original existence proof had $s > 3/2$ but (geometric) uniqueness required $s > 5/2$, see Choquet-Bruhat, Theorem 8.4, p. 168 (note that her s is our $s-1$ so that our $s > \frac{1}{2}n$ is her $s > \frac{1}{2}n + 1$, etc.). For $s > 3/2$, for existence and uniqueness, also in Theorem 7.8 and Theorem 7.10, see Chruściel (2014), Theorem 1.1, or, using very different techniques, Fischer & Marsden (1979), Theorem 4.24. The world record is $s = 1$, i.e. $\tilde{g} \in H^2(\Sigma)$ and $\tilde{k} \in H^1(\Sigma)$ (Klainerman, Rodnianski, & Szeftel, 2015).

This does not follow from (the proof of) Theorem 7.8 alone (i.e. by reduction to a wave gauge). In addition, one needs a uniqueness (or causality) result for quasi-linear wave equations, which states that if two solutions have the same initial data on some submanifold $\Sigma_0 \subset \Sigma$, then they coincide on the domain of dependence $D^+(\Sigma_0)$. See Appendix B.3 for some more background.[339]

In sum, a MGHD $(M, g, \imath)$ of initial data $(\Sigma, \tilde{g}, \tilde{k})$ for the Einstein equations enjoys:[340]

1. ***Existence***, with satisfactory regularity dictated by regularity of the initial data $(\Sigma, \tilde{g}, \tilde{k})$.
2. ***Maximality***, at least within the realm of globally hyperbolic solutions.
3. ***Uniqueness*** up to isometry, in the precise sense stated in Theorem 7.10.
4. ***Causal propagation***, in that initial data at $\Sigma_0 \subset \Sigma$ determine the solution within $D^+(\Sigma_0)$.
5. ***Cauchy stability***, in that the 4-metric g continuously depends on the initial data $(\Sigma, \tilde{g}, \tilde{k})$.

These features of the initial-value approach to GR have given rise to an ideology in which:

- All valid *assumptions* about GR are assumptions about initial data $(\Sigma, \tilde{g}, \tilde{k})$.
- All valid *questions* in GR are questions about the MGHD $(M, g, \imath)$ of these data.

This PDE-based program has so far had spectacular successes.[341] It sometimes gives a slightly different perspective from the (Penrosian) mathematical approach to GR originating in the 1960s, in which typically larger (e.g. analytically extended) space-times are studied. See §§10.4–10.5.

In fact, even the PDE results stated above should be seen as "classical" in the somewhat different sense that they used spacelike Cauchy surfaces. Since the 1990s, much progress in the initial-value approach to GR has been made by giving initial data on certain null hypersurfaces, which lead to a ***characteristic initial value problem***.[342] The idea of solving PDEs through characteristics originally came from first-order PDEs.[343] The simplest version is the PDE

$$\mathscr{L}_X f = 0, \tag{7.168}$$

where $X \in \mathfrak{X}(M)$. This is solved by any $f \in C^\infty(M)$ that is constant along the integral curves (i.e. flow) of the vector field X, which in this context are called the ***characteristics*** of the PDE.

Thus the PDE is effectively replaced by an ODE, namely integrating X. In the usual Cauchy problem, one fixes a solution f by prescribing its value on a non-characteristic (Cauchy) surface $\Sigma \subset M$, in the sense that the characteristics are nowhere tangent to Σ (otherwise, one may have constraints on the initial data and have both an under- and overdetermined problem).

[339] For a more detailed treatment cf. Choquet-Bruhat (2009), Appendix III, Theorem 2.15.

[340] These points are developed in far greater detail in Choquet-Bruhat (2009) and Ringström (2009, 2013).

[341] These started with the proof of stability of Minkowski space-time under small perturbations of the initial data (Christodoulou & Klainerman, 1993), and at the time this book went to press culminated in analogous stability results for the Schwarzschild metric (Dafermos, Holzegel, Rodnianski, & Taylor, 2021) and the slowly rotating Kerr metric (Häfner, Hintz, & Vasy, 2019; Klainerman & Szeftel, 2021). See also the references in footnote 297.

[342] For GR this goes back to Penrose (1963), written in 1961 and republished in 1980, and Bondi and Sachs (see references in Chruściel & Paetz, 2012). In Penrose's spinorial approach (see also Penrose & Rindler, 1984, and more briefly Stewart, 1991) there are no constraints at all. It was further developed by Friedrich (1979), and, for numerical relativity, by Stewart & Friedrich (1982) and Friedrich & Stewart (1983). Existence theorems go back to Rendall (1990) and were later improved by Luk (2012). See also Christodoulou & Klainerman (1993), Klainerman & Nicolò (2003a), Christodoulou (2008), Choquet-Bruhat, Chruściel, & Martín-García (2011), and Aretakis (2013).

[343] For the classical theory see e.g. Courant & Hilbert (1962) or Rauch (2012).

The general idea of solving or at least simplifying some PDE by solving an associated "characteristic" ODE also works for certain second-order hyperbolic PDEs. The simplest example is the wave equation $(-\partial_t^2+\partial_x^2)f=0$ in $d=2$. With $u=t-x$ and $v=t+x$, this is solved by $f(u,v)=g(u)+h(v)$. In other words, any function that is constant along either all characteristics $u=$ constant or along all characteristics $v=$ constant is a solution. In the usual Cauchy problem one gives initial data $f(0,x)$ and $\dot{f}(0,x)$ at $t=0$, or on more general spacelike Cauchy surfaces Σ, since as long as Σ is spacelike the characteristics are nowhere tangent to it. However, in this case it is perfectly reasonable, and perhaps even more natural, to prescribe initial data on some fixed characteristic $u=$ constant together with a fixed characteristic $v=$ constant. For example, one may take the lightcone $(u=0)\cup(v=0)$, which obviously fixes both g and h, and hence f. This also works locally, in the sense that we may take two finitely extended fd lightlike lines N_1 and N_2 that emanate from the same point (or, from a different point of view, would intersect at that point), forming a "V" (the apex is not supposed to be part of either N_1 or N_2). In that case, prescribing f on $N_1\cup N_2$ fixes the solution (still to the $2d$ wave equation) at least on the future domain of dependence $D^+(N_1\cup N_2)$, as one easily verifies from a picture.

This $2d$ setting has two different generalizations to $d=3$ or 4. One may specify initial data:

- either on an open (truncated or semi-infinite) fd null cone emanating from its apex;[344]
- or on two bounded open null hypersurfaces $N_1=C$, $N_2=\underline{C}$ as described in §6.3.

We briefly summarize the latter scheme, which is more popular than the former. In $d=4$, or in $d=3$, where the two-sphere S^2 is replaced by the circle S^1, C ($\underline{C}$) is a null hypersurface that is: (i) bounded in the past by a spacelike sphere; (ii) generated by the fd lightlike geodesics integrating the lightlike vector field L ($\underline{L}$), and (iii) foliated by two-spheres S_t ($\underline{S}_{\underline{t}}$), see (6.61), (6.81), for some range $0<t<t_f$ ($0<\underline{t}<\underline{t}_f$) for which C ($\underline{C}$) is smooth. If (x^1,x^2) are coordinates on S^2, then (x^1,x^2,t) and $(x^1,x^2,\underline{t})$ are coordinates on N_1 and N_2, respectively.

In the wave gauge (7.105), suitable "characteristic" initial data on $N_1\cup N_2$ for the Einstein equations in vacuum (as well as for certain matter sources, including electromagnetism) are provided by a family $t\mapsto\tilde{g}_t$ of $2d$ Riemannian metrics on the spheres S_t foliating C, plus a family $t\mapsto k_t$ of covariant symmetric 2-tensors playing the role of (6.73), i.e. of the "null extrinsic curvature", and similar data on N_2. These are supplemented by a scalar function and a 1-form on S^2. The first of these will be initial value for the $g_{\underline{t}\underline{t}}$ component of the $4d$ metric g, whilst the second is an initial value for what is called the "torsion" $X\mapsto\zeta(X):=g(\nabla_X L,\underline{L})$.

These initial data are constrained in a very different way from the spacelike case. Apart from certain continuity and compatibility requirements, the key constraint on the tensors $\tilde{g}_t$ and k_t on N_1 is given by the null Raychaudhuri equation (6.98), in which θ is *defined* by (6.78), and by a similar equation for the initial data on N_2. These Raychaudhuri equations are ODEs (as opposed to the elliptic PDEs in the usual approach), which is of course a major simplification. In fact, as in the spacelike case (cf. §8.6), but with very different details, unconstrained initial data may be given using conformal methods. In particular, the unconstrained metric data are conformal equivalence classes of such families of $2d$ Riemannian metrics on the spheres S_t and $\underline{S}_{\underline{t}}$.

This leads to a counterpart to Theorem 7.7, i.e. one locally obtains globally hyperbolic solutions from such initial data, which are unique in appropriate coordinates.[345] However, the analogue of a coordinate-free Cauchy development of the initial data remains to be formulated precisely. *A fortiori*, a "characteristic" version of Theorem 7.10 is still waiting to be proved.

[344] See Choquet-Bruhat, Chruściel, & Martín-García (2011) for this.

[345] See Luk (2012), who extended the region in which Rendall (1990) proved the existence of solutions.

7.7 Geometric form of the constraints

In this section we pay our debt by (twice!) deriving the (vacuum) constraints (7.148) - (7.149); matter sources give additional terms in the constraints, see (8.65) - (8.67). Thus we assume a spacelike hypersurface $\Sigma \subset M$ of a space-time (M,g) that solves the vacuum Einstein equations (it is not necessary for this derivation that Σ be a Cauchy surface). The constraints are geometric and hence coordinate-independent, but their derivation is most easily done in coordinates (x^μ) where (x^1,x^2,x^3) are coordinates on Σ, $g_{00} = -1$, and $g_{0i} = 0$. Such coordinates always exist locally, see Proposition 8.1 in §8.1. In such coordinates, the fd unit normal to Σ is simply

$$N^\mu = \partial_0 = (1,0,0,0); \qquad N_\mu = (-1,0,0,0). \tag{7.169}$$

In such coordinates, the Gauss relation (4.148) reads, with spatial indices $i,j,k,l = 1,2,3$,

$$R_{ijkl} = \tilde{R}_{ijkl} + \tilde{k}_{ik}\tilde{k}_{jl} - \tilde{k}_{il}\tilde{k}_{jk}. \tag{7.170}$$

Contracting this to the spatial Ricci tensor $R_{ij} = g^{\mu\nu}R_{\mu i\nu j}$ and Ricci scalar $R = g^{\mu\nu}R_{\mu\nu}$ gives

$$R_{ij} + R_{0i0j} = \tilde{R}_{ij} + \mathrm{Tr}\,(\tilde{k})\tilde{k}_{ij} - \tilde{k}_{ij}^2; \tag{7.171}$$

$$R + 2R_{00} = \tilde{R} + \mathrm{Tr}\,(\tilde{k})^2 - \mathrm{Tr}\,(\tilde{k}^2), \tag{7.172}$$

so that (7.148) is precisely the geometric form of the so-called ***Hamiltonian constraint***

$$G_{00} := R_{00} - \tfrac{1}{2}g_{00}R = R_{00} + \tfrac{1}{2}R = 0. \tag{7.173}$$

In the same sprit, in our coordinate system Codazzi's equation (4.149) system comes down to

$$R_{0ijk} = \tilde{\nabla}_i\tilde{k}_{jk} - \tilde{\nabla}_j\tilde{k}_{ik}. \tag{7.174}$$

Contracting to the Ricci tensor gives

$$R_{0i} = g^{\mu\nu}R_{\mu 0\nu i} = g^{jk}R_{j0ki} = g^{jk}R_{0ijk} = \partial_i\mathrm{Tr}\,(\tilde{k}) - \tilde{\nabla}_j\tilde{k}_i^j. \tag{7.175}$$

Contracting to the Ricci scalar is unnecessary, since the ***momentum constraint*** is simply

$$G_{0i} := R_{0i} - \tfrac{1}{2}g_{0i}R = R_{0i} = 0, \tag{7.176}$$

so that (7.149) follows from (7.175). Note that $\partial_i\mathrm{Tr}\,(\tilde{k}) = \tilde{\nabla}_i\mathrm{Tr}\,(\tilde{k})$, as $\mathrm{Tr}\,(\tilde{k})$ is a scalar.

We now also present a coordinate-free proof of (7.148) - (7.149), via a $4d$-version of the $3d$-objects $\tilde{g}$ and $\tilde{k}$ defined on Σ. These are given in any coordinates by

$$\tilde{g}_{\mu\nu} := g_{\mu\nu} + N_\mu N_\nu; \tag{7.177}$$

$$\tilde{k}_{\mu\nu} := -\tilde{g}^\rho_\mu\tilde{g}^\sigma_\nu\nabla_\rho N_\sigma. \tag{7.178}$$

See (6.11) - (6.11).[346] Note that indices are raised and lowered with g, so that

$$\tilde{g}^\nu_\mu = \delta^\nu_\mu + N_\mu N^\nu. \tag{7.179}$$

[346] Note the minus sign in (7.178) compared to (6.12), which is a consequence of different conventions in fluid mechanics and differential geometry. Many physics texts have a plus in (7.178).

This tensor is also called h^ν_μ. Taken at $x \in M$, it is the matrix of the orthogonal projection operator (4.135). Unlike the original $\tilde{g} \in \mathfrak{X}^{(2,0)}(\Sigma)$, the new $\tilde{g} \in \mathfrak{X}^{(2,0)}(M)$ is defined on any pair of vectors $X,Y \in T_xM$ ($x \in \Sigma$), though the extension is somewhat trivial in that $\tilde{g}(X,N) = 0$ for any $X,Y \in T_xM$, whilst $\tilde{g}(X,Y)$ defined from (7.177) equals the original $\tilde{g}(X,Y)$ defined from (4.124). Hence the ambiguous notation is admissible and it is always clear which $\tilde{g}$ is meant. Likewise for $\tilde{k}$ in (7.178). In terms of the projection π_x, for all $x \in \Sigma$ and $X,Y \in \mathfrak{X}(M)$,

$$\tilde{g}_x(X,Y) = g(\pi_x(X),\pi_x(Y)); \tag{7.180}$$
$$\tilde{k}_x(X,Y) = k(\pi_x(X),\pi_x(Y)). \tag{7.181}$$

The Gauss-Codazzi identities (4.148) - (4.149) are now rewritten as

$$\tilde{g}^\mu_\alpha \tilde{g}^\nu_\beta \tilde{g}^\rho_\gamma \tilde{g}^\sigma_\delta R_{\rho\sigma\mu\nu} = \tilde{R}_{\gamma\delta\alpha\beta} + \tilde{k}_{\gamma\alpha}\tilde{k}_{\delta\beta} - \tilde{k}_{\gamma\beta}\tilde{k}_{\alpha\delta}; \tag{7.182}$$
$$\tilde{g}^\mu_\alpha \tilde{g}^\nu_\beta \tilde{g}^\rho_\gamma N^\sigma R_{\rho\sigma\mu\nu} = \tilde{\nabla}_\beta \tilde{k}_{\alpha\gamma} - \tilde{\nabla}_\alpha \tilde{k}_{\beta\gamma}. \tag{7.183}$$

The corresponding contracted Gauss relations easily follow from (7.182), and are given by

$$\tilde{g}^\mu_\alpha \tilde{g}^\nu_\beta R_{\mu\nu} + \tilde{g}^\sigma_\alpha \tilde{g}^\nu_\beta N^\mu N^\rho R_{\rho\sigma\mu\nu} = \tilde{R}_{\alpha\beta} + \mathrm{Tr}\,(\tilde{k})\tilde{k}_{\alpha\beta} - \tilde{k}^2_{\alpha\beta}; \tag{7.184}$$
$$R + 2N^\mu N^\nu R_{\mu\nu} = \tilde{R} + \mathrm{Tr}\,(\tilde{k})^2 - \mathrm{Tr}\,(\tilde{k}^2), \tag{7.185}$$

where we used the following identity and notations:

$$\tilde{g}^{\alpha\gamma}\tilde{g}^\mu_\alpha \tilde{g}^\rho_\gamma = \tilde{g}^{\rho\mu} = g^{\rho\mu} + N^\rho N^\mu; \tag{7.186}$$
$$\mathrm{Tr}\,(\tilde{k}) = \tilde{k}^\mu_\mu = g^{\mu\nu}\tilde{k}_{\mu\nu} = \tilde{g}^{\mu\nu}\tilde{k}_{\mu\nu}; \tag{7.187}$$
$$\mathrm{Tr}\,(\tilde{k}^2) = \tilde{g}^{\mu\nu}\tilde{k}^2_{\mu\nu} = \tilde{g}^{\mu\nu}\tilde{k}_{\mu\rho}\tilde{k}^\rho_\nu = \tilde{g}^{\mu\nu}\tilde{g}^{\rho\sigma}\tilde{k}_{\mu\rho}\tilde{k}_{\nu\sigma}. \tag{7.188}$$

If we now write the Hamiltonian constraint $G_{00} = 0$ in pseudo-covariant form as

$$N^\mu N^\nu G_{\mu\nu} = N^\mu N^\nu (R_{\mu\nu} - \tfrac{1}{2} g_{\mu\nu} R) = N^\mu N^\nu R_{\mu\nu} + \tfrac{1}{2} R = 0, \tag{7.189}$$

it is clear that (7.185) and (7.189) reproduce (7.148).

Similarly, the contracted Codazzi relations (which stop at one stage) follow from (7.183) as

$$N^\mu \tilde{g}^\nu_\alpha R_{\mu\nu} = \tilde{\partial}_\alpha \mathrm{Tr}\,(\tilde{k}) - \tilde{\nabla}_\mu \tilde{k}^\mu_\alpha. \tag{7.190}$$

The momentum constraint $G_{i0} = 0$ is now written pseudo-covariantly as

$$N^\mu \tilde{g}^\nu_\alpha G_{\mu\nu} = N^\mu \tilde{g}^\nu_\alpha (R_{\mu\nu} - \tfrac{1}{2} g_{\mu\nu} R) = N^\mu \tilde{g}^\nu_\alpha R_{\mu\nu} = 0, \tag{7.191}$$

since $g_{\mu\nu} N^\mu \tilde{g}^\nu_\alpha = 0$. With (7.190), this recovers (7.149), and we are ready. See also §8.3.

The Gauss relation (4.148), or, equivalently, (7.170) or (7.182), describes the value of the Riemann tensor $R(W,Z,X,Y)$ at four spatial vectors (W,Z,X,Y), whereas the Codazzi relation (4.149), or (7.174) or (7.183), gives its value $R(N,Z,X,Y)$ at three spatial directions X,Y,Z and one timelike direction N. For the dynamical (evolution) equations $G_{ij} = 0$ we will also need the case $R(W,N,X,N)$ of two spatial and two orthogonal timelike vectors; unlike the previous two cases, which just rely on the embedding $\Sigma \subset M$, this new case will contain derivatives of $\tilde{g}_{ij}$ and $\tilde{k}_{ik}$ in the orthogonal (temporal) direction, i.e., in suitable coordinates, $\partial_t \tilde{g}_{ij}$ and $\partial_t \tilde{k}_{ik}$. This requires not just a single Cauchy surface $\Sigma \subset M$, but a foliation $M = \sqcup_t \Sigma_t$. This is the subject of the next chapter; the required identity will be (8.37), or, equivalently, (8.38) or (8.39).

8 The 3+1 split of space-time

In this chapter we develop the non-covariant approach of §7.5 through a split of space-time into space and time.[347] Philosophers would say that this split relates the "scientific" image of GR to its "manifest image", since what we experience is space and time separately, rather than Minkowski's (and subsequently also Einstein's) lofty notion of space-time. The $3+1$ split is the key to e.g. the Hamiltonian approach to GR discussed in §8.7, as well as to numerical relativity.

8.1 Lapse and shift

In the previous section we described the constraints $G_{\mu 0} = 0$ in $3+1$ split geometric form (7.148) - (7.149). These constraints do not contain time derivatives of $\tilde{g}_{ij}$ and $\tilde{k}_{ij}$, whose time-evolution is governed by the spatial Einstein equations $G_{ij} = 0$. To rewrite these in $3+1$ form it is not enough to have a single Cauchy surface $\Sigma \subset M$; we need to assume a foliation

$$M = \sqcup_t \Sigma_t \tag{8.1}$$

of M by spacelike Cauchy surfaces Σ_t. In particular, we assume that (M, g) is globally hyperbolic.

The choice of a foliation may be compared with a choice of gauge in the covariant approach in §7.5, like the wave gauge (7.105), whose goal it is to single out a unique metric solving the Einstein equations within its equivalence class under diffeomorphisms. A foliation by spacelike hypersurfaces is a choice of a "now" at each instant of time; it is hallmark of GR that such a choice is arbitrary (as long as each Σ_t is spacelike). See §1.10 and §8.11. As explained in §7.5, given such gauge fixing on all of M, one only needs to solve the *spatial* Einstein equations $G_{ij} = 0$, and impose $G_{\mu 0} = 0$, i.e. (7.148) - (7.149), as constraints on the initial value surface Σ. See §8.3. In the light of Theorem 5.44, such a foliation is equivalent to a diffeomorphism

$$F : \mathbb{R} \times \Sigma \to M \tag{8.2}$$

with the property that each subspace $\Sigma_t := F_t(\Sigma)$ is spacelike. With $t \in \mathbb{R}$ and $x \in \Sigma$, we write

$$F_t : \Sigma \to M; \qquad F_t(x) := F(t,x); \tag{8.3}$$

$$F_x : \mathbb{R} \to M; \qquad F_x(t) := F(t,x), \tag{8.4}$$

which shows the double role of foliations: for fixed time $t \in \mathbb{R}$ the map F_t is a spacelike embedding of Σ in M, whereas for fixed $x \in \Sigma$ the map F_x is a curve through $F(0,x) \in M$. *A priori* defined by F, such a foliation (8.1) is also equivalent to one of the following structures:

- A temporal function $\mathrm{t} : M \to \mathbb{R}$ with $g(\nabla \mathrm{t}, \nabla \mathrm{t}) < 0$, cf. Definition 5.42 and Theorem 5.44.
- A function called the ***lapse*** L and a vector field called the ***shift*** S of the foliation.[348]

The lapse and shift may be defined in a coordinate-independent way by the decomposition

$$\frac{dF_t}{dt} =: LN + S, \tag{8.5}$$

seen as an equality between vectors in $T_y M$ for any $y = F(x,t)$, as follows:[349]

[347] The $3+1$ split originated in the work of Darmois (1927), Lichnerowicz (1939, 1955), and Fourès-Bruhat (1956); see Choquet-Bruhat (2018). It subsequently crossed the independent development of the Hamiltonian formalism for GR, in particular through the work of Arnowitt, Deser, & Misner (1962). See footnote 384 in §8.7.

[348] One may wonder why something can simultaneously be determined by one function t and by four functions L and $\vec{S}$, but the metric information in the former is in the four components of the vector field $\nabla \mathrm{t}$.

[349] Many authors write (8.5) as $\partial_t = Nn + \vec{N}$, where N is the lapse, n is the normal, and $\vec{N}$ is the shift.

- the left-hand side is the tangent vector at y to (e.g.) the curve $c(s) = F_x(t+s)$ at $s = 0$;
- $L \in C^\infty(M)$ is a scalar whilst $N \in \mathfrak{X}(M)$ is the normal future-directed vector field to Σ_t;
- $S \in \mathfrak{X}(M)$ is the orthogonal projection of dF_t/dt onto $T_y\Sigma_t \subset T_tM$, hence tangent to Σ_t.

Here we assume (4.131). Thus, given the metric g on M, a foliation F of M by spacelike Cauchy surfaces uniquely defines L and S. Conversely, the idea is that L and S fix a foliation (8.1), but not all pairs (L,S) do so, not even if $L > 0$. Starting from a Cauchy surface $\Sigma \subset M$ it turns out that one may always *globally* put $S = 0$, see (8.14) below and Theorem 5.44 (although this may not be the wisest choice). In addition, one may *locally* set $L = 1$ (see Proposition 8.1 below), but the latter is generally not possible globally: if $S = 0$ and $L = 1$, then the flow lines of N would be (pre)geodesics, whose focusing and hence crossing (in the presence of positive curvature) would invalidate the foliation. There might be similar problems with other choices of L and S.

From the point of view of a temporal function t, the lapse and shift are given by

$$L = \frac{1}{\sqrt{-g(\nabla\mathsf{t},\nabla\mathsf{t})}}; \qquad N = -L\nabla\mathsf{t}. \tag{8.6}$$

We can choose coordinates (x^0,x^1,x^2,x^3) *adapted to the foliation* (8.1), as follows:

- $x^0 = t$, or, more precisely, $x^0(x) = t$, provided $x \in \Sigma_t$;
- (x^i) are (local) coordinates initially on Σ ($i = 1,2,3$), but subsequently on any slice Σ_t: if $y \in \Sigma_t$, the flow line of the vector field $\nabla\mathsf{t}$ (or N) hits Σ in exactly one point $x_0 \in \Sigma$; if the latter has coordinates $x_0 = (0,x^1,x^2,x^3)$, the former has coordinates $y = (t,x^1,x^2,x^3)$.

Given (local) spatial coordinates (x^1,x^2,x^3) on Σ, at any point $x \in \Sigma_t$ one has tangent vectors $e_i = \partial_i$ to Σ_t, as well as a one-form $\theta^0 = d\mathsf{t}$. As we have seen, ∂_0 may not be orthogonal to Σ_t and hence to the vectors e_i, but the shift $S = S^i\partial_i := \sum_{i=1}^3 S^i\partial_i$ corrects for this, in that the vector

$$e_0 = \partial_0 - S \tag{8.7}$$

is orthogonal to Σ. We then have a frame (e_a) with dual coframe (θ^b), defined by

$$e_0 := \partial_t - S^i\partial_i; \qquad e_i := \partial_i; \tag{8.8}$$

$$\theta^0 := dt; \qquad \theta^i := dx^i + S^i dt, \tag{8.9}$$

where $g(e_0,e_i) = 0$ and $g(\theta^0,\theta^i) = 0$, and, by definition, $\theta^a(e_b) = \delta^a_b$ for $a,b = 0,1,2,3$.

By definition of the lapse and the shift, we then have the useful relations

$$g = -L^2(\theta^0)^2 + \tilde{g}_{ij}\theta^i\theta^j; \qquad e_0 = LN = -L^2\nabla\mathsf{t}; \tag{8.10}$$

$$d\mathsf{t} = dt; \qquad \nabla\mathsf{t} = g^{\mu 0}\partial_\mu; \tag{8.11}$$

$$L = 1/\sqrt{-g^{00}}; \qquad S^i = -g^{i0}/g^{00}; \tag{8.12}$$

$$N_\mu = (-L,0,0,0); \qquad N^\mu = (1/L,-S/L). \tag{8.13}$$

Consequently, in coordinates adapted to the foliation, the metric and its inverse take the form

$$g_{\mu\nu} = \begin{pmatrix} -L^2 + S_jS^j & S_i \\ S_i & \tilde{g}_{ij} \end{pmatrix}; \qquad g^{\mu\nu} = \begin{pmatrix} -1/L^2 & S^i/L^2 \\ S^i/L^2 & \tilde{g}^{ij} - S^iS^j/L^2 \end{pmatrix}, \tag{8.14}$$

where $\tilde{g}^{ij}$ is the matrix inverse to $\tilde{g}_{ij}$ and spatial indices are raised and lowered with this spatial metric (so that e.g. $S_jS^j = \tilde{g}_{ij}S^iS^j$). Thus L and S^i may also be seen as parametrizations of the non-spatial components of the metric. The local possibilities are as follows:[350]

Proposition 8.1 *For any Cauchy surface $\Sigma \subset M$ with 3-metric $\tilde{g}_{ij}$ and extrinsic curvature $\tilde{k}_{ij}$ in given coordinates (x^μ), there exist coordinates (y^μ) in which $\tilde{g}_{ij}$ and $\tilde{k}_{ij}$ are* the same*, whilst*

$$g^{i0} = g_{0i} = 0; \qquad g^{00} = g_{00} = -1 \tag{8.15}$$

on Σ. Moreover, in a nbhd of Σ one can give the components $g_{0\mu}$ any desired value.

We now expand the pseudo-covariant notation (7.177) - (7.178), originally defined on $\Sigma \subset M$, to all of M, assuming (8.1) and the ensuing extension of the normal vector field from Σ to M. Then

$$k_{\mu\nu} := -\nabla_\mu N_\nu = \tilde{k}_{\mu\nu} + N_\mu A_\nu, \tag{8.16}$$

where the ***acceleration*** A of the vector field N is defined by

$$A = \nabla_N N; \qquad A^\mu = N^\nu \nabla_\nu N^\mu. \tag{8.17}$$

We now shed interesting new light on the extrinsic curvature $\tilde{k}$ of $\Sigma \subset M$ by showing that

$$\tilde{k} = -\tfrac{1}{2}\mathscr{L}_N \tilde{g} \tag{8.18}$$
$$= -\tfrac{1}{2}L^{-1}\mathscr{L}_{e_0}\tilde{g}, \tag{8.19}$$

seen as equalities between symmetric tensors in either $\mathfrak{X}^{(2,0)}(\Sigma)$ or $\mathfrak{X}^{(2,0)}(M)$; in the former case the proof of (8.18) in fact implies that $\mathscr{L}_N\tilde{g} \in \mathfrak{X}^{(2,0)}(\Sigma)$. In arbitrary coordinates, we have

$$\tilde{k}_{\mu\nu} = -\tfrac{1}{2}\mathscr{L}_N \tilde{g}_{\mu\nu}, \tag{8.20}$$
$$= -\tfrac{1}{2}L^{-1}\mathscr{L}_{e_0}\tilde{g}_{\mu\nu}. \tag{8.21}$$

In coordinates (t, x^i) we may restrict to spatial indices: using (8.8) and (2.94), eq. (8.21) is

$$(\partial_t - \mathscr{L}_S)\tilde{g}_{ij} = -2L\tilde{k}_{ij}, \tag{8.22}$$

which in coordinates where also $L = 1$ and $S = 0$ further simplifies to the transparent equality

$$\tilde{k}_{ij} = -\tfrac{1}{2}\partial_t \tilde{g}_{ij}. \tag{8.23}$$

Before embarking on the the derivation of (8.18) - (8.19), note that (8.23) is easy to derive:

$$\tilde{k}_{ij} = -\nabla_i N_j = -\partial_i N_j + \Gamma^\mu_{ij} N_\mu = -\Gamma^0_{ij} = \tfrac{1}{2}g^{00}\partial_t g_{ij} = -\tfrac{1}{2}\partial_t \tilde{g}_{ij}, \tag{8.24}$$

since in coordinates where $L = 1$ and $S = 0$ we have (8.15) and hence (7.169), cf. (8.14).

To derive (8.18), we first use the $(1,0)$ case of (3.72) with $X = N$ to compute

$$\mathscr{L}_N N_\mu = N^\nu \nabla_\nu N_\mu + (\nabla_\mu N^\nu) N_\nu = N^\nu \nabla_\nu N_\mu = \nabla_N N_\mu, \tag{8.25}$$

[350] This proposition is slightly adapted from Chruściel (2010), Proposition 1.4.1, which is also proved there.

since in the second term $(\nabla_\mu N^\nu)N_\nu$ vanishes because of (4.131), which gives

$$N^\nu \nabla_\mu N_\nu = g(N, \nabla_\mu N) = \tfrac{1}{2}\partial_\mu g(N,N) = \tfrac{1}{2}\partial_\mu(-1) = 0. \tag{8.26}$$

Using this as well as (8.16), the $(2,0)$ case of (3.72) with $X = N$ then gives

$$\mathscr{L}_N(N_\mu N_\nu) = N_\mu \nabla_N N_\nu + N_\nu \nabla_N N_\mu = N_\mu A_\nu + N_\nu A_\mu. \tag{8.27}$$

From (7.177), (3.73), (8.16), and (8.27) we then obtain, at last,

$$\begin{aligned} \mathscr{L}_N \tilde{g}_{\mu\nu} &= \mathscr{L}_N(g_{\mu\nu} + N_\mu N_\nu) = -2\tilde{k}_{\mu\nu} - N_\mu A_\nu - N_\nu A_\mu + N_\mu A_\nu + N_\nu A_\mu \\ &= -2\tilde{k}_{\mu\nu}. \end{aligned} \tag{8.28}$$

We derive (8.19) from (8.18) using a general fact, namely, using (8.17),

$$A_\mu = \tilde{\partial}_\mu(\ln L) = L^{-1}\tilde{g}^\nu_\mu \partial_\nu L, \tag{8.29}$$

where we use the notation $\tilde{\partial}_\mu = \tilde{g}^\nu_\mu \partial_\nu$ for the derivative along Σ.[351] Note that the projection $\tilde{g}^\nu_\mu$ reconfirms that A is tangent to Σ (i.e., orthogonal to N), which we already knew because

$$g(N, \nabla_N N) = 0. \tag{8.30}$$

Using (8.16) and (7.177), eq. (8.29) is equivalent to

$$\nabla_N N_\nu \equiv N^\mu \nabla_\mu N_\nu = L^{-1}(N^\mu N_\nu \partial_\mu + \partial_\nu)L, \tag{8.31}$$

which we will now prove. The proof relies on torsion-freeness of ∇, which implies $\nabla_\mu \partial_\nu f = \nabla_\nu \partial_\mu f$ for any $f \in C^\infty(M)$. We write (8.13) as $N_\mu = -L\partial_\mu t$ and compute

$$\begin{aligned} N^\mu \nabla_\mu N_\nu &= -N^\mu \nabla_\mu (L\partial_\nu t) \\ &= -N^\mu(\partial_\mu L \partial_\nu t + L\nabla_\nu \partial_\mu t) \\ &= L^{-1}N^\mu N_\nu \partial_\mu L - LN^\mu \nabla_\nu (L^{-1}N_\mu) \\ &= L^{-1}N^\mu N_\nu \partial_\mu L - N^\mu N_\mu \partial_\nu L^{-1} - LN^\mu \nabla_\nu N_\mu \\ &= L^{-1}(N^\mu N_\nu \partial_\mu + \partial_\nu)L, \end{aligned} \tag{8.32}$$

where we used (8.26). Using (8.10), (2.94), and (8.18), we then compute

$$\begin{aligned} \mathscr{L}_{e_0}\tilde{g}_{\mu\nu} &= \mathscr{L}_{LN}\tilde{g}_{\mu\nu} \\ &= L\mathscr{L}_N \tilde{g}_{\mu\nu} + N_\mu \partial_\nu L + N_\nu \partial_\mu L + (\partial_\mu L)N^\rho N_\rho N_\nu + (\partial_\nu L)N^\rho N_\rho N_\mu \\ &= L\mathscr{L}_N \tilde{g}_{\mu\nu} + N_\mu \partial_\nu L + N_\nu \partial_\mu L - N_\nu \partial_\mu L - N_\mu \partial_\nu L \\ &= -2L\tilde{k}_{\mu\nu}. \end{aligned} \tag{8.33}$$

This exemplifies a general phenomenon concerning $\mathscr{L}_{e_0}$: if any tensor $\tau \in \mathfrak{X}^{(k,0)}(M)$ satisfies

$$\tau(X_1, \ldots, X_k) = \tau(\pi(X_1), \ldots, \pi(X_k)), \tag{8.34}$$

i.e., τ is purely spatial, or, equivalently $\tau(X_1,\ldots,X_k) = 0$ if $X_i = N$ for at least one i, then also

$$\mathscr{L}_{e_0}\tau(X_1, \ldots, X_k) = \mathscr{L}_{e_0}\tau(\pi(X_1), \ldots, \pi(X_k)), \tag{8.35}$$

that is, also $\mathscr{L}_{e_0}\tau$ is purely spatial. This most easily follows from the Leibniz rule for $\mathscr{L}$ and hence the case $k = 1$. Since $e_0 = LN$ we may as well derive $(\mathscr{L}_{e_0}\tau)(e_0) = 0$ from the assumption $\tau_{e_0}(e_0) = 0$: using (2.94) and $\mathscr{L}_{e_0}e_0 = [e_0, e_0] = 0$, we obtain

$$(\mathscr{L}_{e_0}\tau)(e_0) = e_0(\tau(e_0)) + \tau(\mathscr{L}_{e_0}e_0) = 0 + 0 = 0. \tag{8.36}$$

[351] This is consistent with notation $\tilde{\nabla}$ for the covariant derivative within Σ defined with respect to $\tilde{g}$ because of (4.136), which in coordinates reads $\tilde{g}^\nu_\mu \nabla_\nu Y^\rho = \tilde{\nabla}_\mu Y^\rho$.

8.2 Beyond Gauss-Codazzi: The Darmois identity

As promised at the end of §7.7, we now derive an identity for $\mathrm{Riem}(W,N,X,N)$, the Riemann tensor at two spatial and two orthogonal timelike vectors. This is the final identity in a chain:

- the first such identity was (4.147) - (4.148), with zero entries of N (due to Gauss);
- the second was (4.149), with one slot occupied by N (due to Codazzi);
- the third will be (8.37) below, involving two copies of N (due to Darmois).

More N's are fruitless, as the Riemann tensor vanishes due to its (anti)symmetries. This new case will contain expressions like $\mathscr{L}_{e_0}\tilde{k}$, which unlike terms like $\tilde{\nabla}_l\tilde{k}_{ik}$ in (4.149) involves derivatives in the orthogonal direction. Thus the case of two orthogonal vectors relies on the time function, or, equivalently, on the foliation (8.1) (at least near $\Sigma\equiv\Sigma_0$). The ***Darmois identity***, then, reads

$$\mathrm{Riem}(W,N,X,N) = L^{-1}(\mathscr{L}_{e_0}\tilde{k}(X,W) + \tilde{\nabla}_W\tilde{\nabla}_X L) + \tilde{k}^2(X,W), \tag{8.37}$$

where $X,W\in Tx\Sigma$. In general coordinates, this expression reads

$$\tilde{g}^\rho_\alpha\tilde{g}^\mu_\beta N^\sigma N^\nu R_{\rho\sigma\mu\nu} = L^{-1}(\mathscr{L}_{e_0}\tilde{k}_{\alpha\beta} + \tilde{\nabla}_\alpha\tilde{\nabla}_\beta L) + \tilde{k}^2_{\alpha\beta}, \tag{8.38}$$

where $\tilde{k}^2_{\alpha\beta}\equiv\tilde{k}_{\alpha\rho}\tilde{k}^\rho_\beta$, in which the indices on $\tilde{k}$ are raised and lowered with either $\tilde{g}$ or g (this does not matter because any action of the terms $N_\mu N_\nu$ in (7.177) contracts to zero on $\tilde{k}$), and $\tilde{\nabla}_\beta L = \tilde{\partial}_\beta L$. In coordinates (t,x^i) with zero shift and unit lapse, as before, eq. (8.37) is simply

$$R_{i0j0} = \partial_t\tilde{k}_{ij} + \tilde{k}^2_{ij}. \tag{8.39}$$

To see this, eq. (4.13) gives $R_{i0j0} = (\nabla_j\nabla_0 - \nabla_0\nabla_j)N_i$. Eqs. (8.15) and (7.169) then give

$$\begin{aligned}
&\nabla_j\nabla_0 N_i = \partial_j\nabla_0 N_i - \Gamma^\nu_{0j}\nabla_\nu N_i - \Gamma^\nu_{ij}\nabla_0 N_\nu = -\Gamma^k_{0j}\nabla_k N_i - \Gamma^0_{ij}\nabla_0 N_0 = \Gamma^k_{0j}\tilde{k}_{ki} - \Gamma^0_{ij}\Gamma^0_{00} = -\tilde{k}^2_{ij};\\
&-\nabla_0\nabla_j N_i = \nabla_0\tilde{k}_{ij} = \partial_0\tilde{k}_{ij} - \Gamma^l_{0i}\tilde{k}_{lj} - \Gamma^l_{0j}\tilde{k}_{li} = \partial_0\tilde{k}_{ij} + \tilde{k}^l_i\tilde{k}_{lj} + \tilde{k}^l_j\tilde{k}_{li} = \partial_t\tilde{k}_{ij} + 2\tilde{k}^2_{ij},
\end{aligned}$$

since $\nabla_0 N_i = -\Gamma^\mu_{0i}N_\mu = \Gamma^0_{0i} = 0$, $\Gamma^k_{0j} = \frac{1}{2}g^{kl}\partial_0 g_{jl} = -\tilde{k}^k_j$ from (8.23), $\Gamma^0_{00}=0$, and $\partial_0\equiv\partial_t$.

To derive the coordinate-free version (8.38), we first note that (8.16) and (8.29) give

$$\nabla_\mu N_\nu = -\tilde{k}_{\mu\nu} - N_\mu\tilde{\partial}_\nu(\ln L). \tag{8.40}$$

As in the derivation of the Gauss–Codazzi equations, we start from (4.13), this time with $Z=N$:

$$\begin{aligned}
R^\rho_{\sigma\mu\nu}N^\sigma &= (\nabla_\mu\nabla_\nu - \nabla_\nu\nabla_\mu)N^\rho = -\nabla_\mu(\tilde{k}^\rho_\nu + N_\nu\tilde{\partial}^\rho L) + \nabla_\nu(\tilde{k}^\rho_\mu + N_\mu\tilde{\partial}^\rho L)\\
&= \nabla_\nu\tilde{k}^\rho_\mu - \nabla_\mu\tilde{k}^\rho_\nu + (\nabla_\nu N_\mu - \nabla_\mu N_\nu)\tilde{\partial}^\rho L + (N_\mu\nabla_\nu - N_\nu\nabla_\mu)\tilde{\partial}^\rho L.
\end{aligned} \tag{8.41}$$

This gives

$$N^\sigma N^\nu R_{\rho\sigma\mu\nu} = \nabla_N\tilde{k}_{\rho\mu} - N^\nu\nabla_\mu\tilde{k}_{\rho\nu} + \tilde{\partial}_\mu(\ln L)\tilde{\partial}_\rho(\ln L) + \nabla_\mu\tilde{\partial}_\rho L + N_\mu\nabla_N\tilde{\partial}_\rho L, \tag{8.42}$$

whose last term will vanish upon contraction with $\tilde{g}^\mu_\beta$ in (8.38). We rewrite the second term $N^\nu\nabla_\mu\tilde{k}_{\rho\nu}$ using the fact that $N^\nu\tilde{k}_{\rho\nu}=0$ and hence also $\nabla_\mu(N^\nu\tilde{k}_{\rho\nu})=0$. This gives

$$-N^\nu\nabla_\mu\tilde{k}^\rho_\nu = \tilde{k}^\rho_\nu\nabla_\mu N^\nu = -\tilde{k}^\rho_\nu\tilde{k}^\nu_\mu - \tilde{k}^\rho_\nu N_\mu\tilde{\partial}^\nu(\ln L), \tag{8.43}$$

whose last term will disappear upon contraction with $\tilde{g}^{\mu}_{\beta}$ in (8.38). We now replace the covariant derivative in the first term $\nabla_N \tilde{k}_{\rho\mu}$ by a Lie derivative. Our favorite rule (3.72) gives

$$\mathscr{L}_{e_0}\tilde{k}_{\rho\mu} = \nabla_{e_0}\tilde{k}_{\rho\mu} + (\nabla_\mu e_0^\nu)\tilde{k}_{\rho\nu} + (\nabla_\rho e_0^\nu)\tilde{k}_{\mu\nu}, \tag{8.44}$$

in right-hand side of which we substitute $e_0 = LN$, and hence

$$\nabla_{e_0} = L\nabla_N. \tag{8.45}$$

Recall that unlike the Lie derivative $\mathscr{L}_X$, the covariant derivative ∇_X is $C^\infty(M)$-linear in X. In the remaining terms we use (8.40). Many of the ensuing terms drop out after contraction with $\tilde{g}^{\rho}_{\alpha}\tilde{g}^{\mu}_{\beta}$, and after a lengthy but straightforward computation we obtain

$$\tilde{g}^{\rho}_{\alpha}\tilde{g}^{\mu}_{\beta}\nabla_N\tilde{k}_{\rho\mu} = L^{-1}\nabla_{e_0}\tilde{k}_{\alpha\beta} + 2\tilde{k}^2_{\alpha\beta}. \tag{8.46}$$

Using (8.44) and (8.46) in (8.42) finally gives (8.37), as follows:

$$\begin{aligned}\tilde{g}^{\mu}_{\alpha}\tilde{g}^{\nu}_{\beta}N^\sigma N^\nu R_{\rho\sigma\mu\nu} &= L^{-1}\mathscr{L}_{e_0}\tilde{k}_{\alpha\beta} + 2\tilde{k}^2_{\alpha\beta} - \tilde{k}^2_{\alpha\beta} + \tilde{\partial}_\alpha(\ln L)\tilde{\partial}_\beta(\ln L) + \tilde{\nabla}_\alpha\tilde{\partial}_\beta L \\ &= L^{-1}(\mathscr{L}_{e_0}\tilde{k}_{\alpha\beta} + \tilde{\nabla}_\alpha\tilde{\nabla}_\beta L) + \tilde{k}^2_{\alpha\beta}.\end{aligned} \tag{8.47}$$

For the Einstein equations we do not need the full Riemann tensor $R_{\rho\sigma\mu\nu}$ but its contractions

$$R_{\mu\nu} := R^{\rho}_{\mu\rho\nu} = g^{\rho\sigma}R_{\rho\mu\sigma\nu}; \tag{8.48}$$

$$R := g^{\mu\nu}R_{\mu\nu}, \tag{8.49}$$

defining the Ricci tensor) and Ricci scalar, respectively. For later use we therefore compute the contractions of (8.38), which are slightly involved. First, (7.184) and (8.38) give

$$\tilde{R}_{\alpha\beta} + \mathrm{Tr}\,(\tilde{k})\tilde{k}_{\alpha\beta} - \tilde{k}^2_{\alpha\beta} - \tilde{g}^{\mu}_{\alpha}\tilde{g}^{\nu}_{\beta}R_{\mu\nu} = L^{-1}(\mathscr{L}_{e_0}\tilde{k}_{\alpha\beta} + \tilde{\nabla}_\alpha\tilde{\nabla}_\beta L) + \tilde{k}^2_{\alpha\beta}, \tag{8.50}$$

from which we obtain

$$\tilde{g}^{\mu}_{\alpha}\tilde{g}^{\nu}_{\beta}R_{\mu\nu} = -L^{-1}(\mathscr{L}_{e_0}\tilde{k}_{\alpha\beta} + \tilde{\nabla}_\alpha\tilde{\nabla}_\beta L) + \tilde{R}_{\alpha\beta} + \mathrm{Tr}\,(\tilde{k})\tilde{k}_{\alpha\beta} - 2\tilde{k}^2_{\alpha\beta}. \tag{8.51}$$

Contracting both sides with $\tilde{g}^{\alpha\beta}$, and defining the $3d$ covariant Laplacian

$$\tilde{\Delta} := \tilde{g}^{\alpha\beta}\tilde{\nabla}_\alpha\tilde{\nabla}_\beta, \tag{8.52}$$

gives

$$R + N^\mu N^\nu R_{\mu\nu} = -L^{-1}(\tilde{g}^{\alpha\beta}\mathscr{L}_{e_0}\tilde{k}_{\alpha\beta} + \tilde{\Delta}L) + \tilde{R} + \mathrm{Tr}\,(\tilde{k})^2 - 2\mathrm{Tr}\,(\tilde{k}^2), \tag{8.53}$$

Since

$$\mathscr{L}_{e_0}\tilde{g}_{\alpha\beta} = -2L\tilde{k}_{\alpha\beta} \tag{8.54}$$

by (8.21), we have

$$\mathscr{L}_{e_0}\tilde{g}^{\alpha\beta} = 2L\tilde{k}^{\alpha\beta}, \tag{8.55}$$

cf. (7.29), and hence

$$\tilde{g}^{\alpha\beta}\mathscr{L}_{e_0}\tilde{k}_{\alpha\beta} = \mathscr{L}_{e_0}\mathrm{Tr}\,(\tilde{k}) - \tilde{k}_{\alpha\beta}\mathscr{L}_{e_0}\tilde{g}^{\alpha\beta} = \mathscr{L}_{e_0}\mathrm{Tr}\,(\tilde{k}) - 2L\mathrm{Tr}\,(\tilde{k}^2), \tag{8.56}$$

where of course $\mathscr{L}_{e_0}\mathrm{Tr}\,(\tilde{k}) = e_0(\mathrm{Tr}\,(\tilde{k}))$. Hence (8.53) may be rewritten as

$$R + N^\mu N^\nu R_{\mu\nu} = -L^{-1}(\mathscr{L}_{e_0}\mathrm{Tr}\,(\tilde{k}) + \tilde{\Delta}L) + \tilde{R} + \mathrm{Tr}\,(\tilde{k})^2. \tag{8.57}$$

Using (7.185), we finally obtain the twice contracted version of (8.38), namely

$$R = \tilde{R} - 2L^{-1}(\mathscr{L}_{e_0}\mathrm{Tr}\,(\tilde{k}) + \tilde{\Delta}L) + \mathrm{Tr}\,(\tilde{k})^2 + \mathrm{Tr}\,(\tilde{k}^2). \tag{8.58}$$

8.3 The 3+1 decomposition of the Einstein equations

We now have all information for projecting the Einstein equations (7.1), with $T_{\mu\nu}$ decomposed according to (7.61), in three different directions, namely, contracting with:[352]

- The *spatial* projection $\tilde{g}^\mu_\alpha \tilde{g}^\nu_\beta$, which gives the ***dynamical equations***

$$\mathscr{L}_{e_0}\tilde{k}_{\mu\nu} = -\tilde{\nabla}_\mu\tilde{\nabla}_\nu L + L(\tilde{R}_{\mu\nu} + \mathrm{Tr}\,(\tilde{k})\tilde{k}_{\mu\nu} - 2\tilde{k}^2_{\mu\nu} + 4\pi((S-E)\tilde{g}_{\mu\nu} - 2S_{\mu\nu})); \quad (8.59)$$

$$\mathscr{L}_{e_0}\tilde{g}_{\mu\nu} = -2L\tilde{k}_{\mu\nu}. \quad (8.60)$$

  These follow from (7.62), (8.51), (7.63), and (8.21). As already noted, in Σ-adapted coordinates eq. (8.60) becomes (8.22), and with (8.59), one may write this system as

$$(\partial_t - \mathscr{L}_S)\tilde{k}_{ij} = -\tilde{\nabla}_i\tilde{\nabla}_j L + L(\tilde{R}_{ij} + \mathrm{Tr}\,(\tilde{k})\tilde{k}_{ij} - 2\tilde{k}^2_{ij} + 4\pi((S-E)\tilde{g}_{ij} - 2S_{ij})); \quad (8.61)$$

$$(\partial_t - \mathscr{L}_S)\tilde{g}_{ij} = -2L\tilde{k}_{ij}, \quad (8.62)$$

  where, using (2.94) and (3.73), respectively, the two Lie derivatives may be written as

$$\mathscr{L}_S\tilde{k}_{ij} = S^l\partial_l\tilde{k}_{ij} + \tilde{k}_{jl}\partial_i S^l + \tilde{k}_{il}\partial_j S^l; \quad (8.63)$$

$$\mathscr{L}_S\tilde{g}_{ij} = \tilde{\nabla}_i S_j + \tilde{\nabla}_j S_i. \quad (8.64)$$

- The *timelike* projections $N^\mu N^\nu$, which gives the ***Hamiltonian constraint***

$$\tilde{R} - \mathrm{Tr}\,(\tilde{k}^2) + \mathrm{Tr}\,(\tilde{k})^2 = 16\pi E, \quad (8.65)$$

  which follows from (7.1) and (7.185). It plays a key role in (canonical) quantum gravity.

- The *mixed* projections $\tilde{g}^\mu_\alpha N^\nu$ or $\tilde{g}^\nu_\beta N^\mu$, producing the ***momentum constraint***

$$\tilde{\nabla}_\mu\tilde{k}^\mu_\nu - \tilde{\nabla}_\nu\mathrm{Tr}\,(\tilde{k}) = 8\pi P_\nu. \quad (8.66)$$

  This follows from (7.1), whose $g_{\mu\nu}R$ term contracts to zero, and (7.190). Equivalently,

$$\tilde{\nabla}_j\tilde{k}^j_i - \tilde{\nabla}_i\mathrm{Tr}\,(\tilde{k}) = 8\pi P_i. \quad (8.67)$$

Altogether, in adapted coordinates, eqs. (8.61), (8.62), (8.65), and (8.67) form a coupled system of 16 PDEs for 16 unknown functions $(\tilde{g}_{ij}, \tilde{k}_{ij}, L, S^i)$ defined on the Cauchy (hyper)surface Σ, where the $\tilde{k}_{ij}$ may be exchanged for the time-derivaties $\partial_t\tilde{g}_{ij}$ through (8.62), leaving 10 coupled PDEs for 10 unknowns $(\tilde{g}_{ij}, L, S^i)$, similar to the original covariant Einstein equations (which are 10 coupled PDEs for the 10 components $g_{\mu\nu}$ of the four-dimensional metric). In the latter case, the spatial part consists of *six evolution equations*, whereas the other two parts contain only first time derivatives of the spatial metric and no time derivates of the lapse and shift functions at all; hence these act as *four constraints* on the initial data $(\tilde{g}_{ij}, \partial_t\tilde{g}_{ij})$, or, in general, on $(\tilde{g}_{ij}, \tilde{k}_{ij})$. Also cf. §7.5. The lapse and shift functions are not determined by the equations at all and hence can be (more or less) freely chosen; doing so amounts to fixing a (local) gauge, see §8.1. In that respect, the diffeomorphism invariance of the original (covariant) Einstein equations (7.1) has been traded for the arbitrariness of the lapse L and the shift $\vec{S}$ and hence of the foliation.

The precise way these equations are equivalent to the Einstein equations is as follows:[353]

[352] The letters S and $S_{\mu\nu}$ on the right-hand sides below refer to the energy-momentum tensor, whereas the S in $\mathscr{L}_S$ on the left and the S^i on the right refer to the shift vector, sorry!

[353] See Fischer & Marsden (1979), Theorem 4.1.

Theorem 8.2 *Let (M,g) be a globally hyperbolic space-time equipped with a foliation (8.1) by spacelike Cauchy surfaces Σ_t, and associated lapse L and shift S. Let $(\tilde{g}(t),\tilde{k}(t))$ be the (Riemannian) 3-metric and exterior curvature on Σ_t induced by the (Lorentzian) 4-metric g. Then g is a solution of the Einstein equations (7.1), possibly coupled to matter with conserved energy-momentum tensor $T_{\mu\nu}$ in the sense that $\nabla^\mu T_{\mu\nu}=0$ holds identically without (7.1),[354] iff:*

1. *For **some** t the pair $(\tilde{g}_{ij}(t),\tilde{k}_{ij}(t))$ satisfies the constraint equations (8.65) and (8.67);*
2. *The maps $t\mapsto\tilde{g}_{ij}(t)$ and $t\mapsto\tilde{k}_{ij}(t)$ satisfy the evolution equations (8.61) - (8.62).*

This follows from our computations showing that (8.61), (8.62), (8.65), and (8.67) are equivalent to the Einstein equations (7.1). Furthermore, the proof in §7.5 that the constraints propagate is the same as for the vacuum case, see (7.126) - (7.127) and surrounding text. □

Conversely, for given lapse $L>0$ and shift S one can only expect existence and uniqueness of an ensuing space-time (M,g) solving the vacuum Einstein equations locally in time, i.e. in some nbdh of Σ, since there is no *a priori* global control over the foliation that (L,S) give rise to.

Theorem 8.2 understates the importance of the constraints for GR. In fact:[355]

Theorem 8.3 *A (globally hyperbolic) space-time (M,g) satisfies the Einstein equations $G_{\mu\nu}=0$ iff the Hamiltonian constraint (7.148) holds on every spacelike (Cauchy) surface $\Sigma\subset M$.*

Proof. The implication from left to right is obvious, so assume (7.148) holds on every spacelike surface (we leave it to the reader to insert the words between brackets in the proof). As we have seen via eqs. (7.189) and (7.185), the Hamiltonian constraint (7.148) is, pseudo-covariantly,

$$N^\mu N^\nu G_{\mu\nu}=0, \tag{8.68}$$

at each $x\in\Sigma$, where N is the (fd) normal to Σ. Requiring this for all spacelike (Cauchy) surfaces Σ comes down to asking (8.68) for every timelike vector field N. If N_1 and N_2 are fd timelike, then so is N_1+N_2, which shows that (8.68) implies the seemingly stronger condition

$$N_1^\mu N_2^\nu G_{\mu\nu}=0, \tag{8.69}$$

for all timelike N_1 and N_2. Furthermore, any spacelike vector X equals $X=N_1-N_2$ for some timelike N_1 and N_2, so that (8.68) implies $N^\mu X^\nu G_{\mu\nu}=0$ for all timelike N and spacelike X, and by the same argument, $X_1^\mu X_2^\nu G_{\mu\nu}=0$ for all spacelike X_1 and X_2. Finally, any vector Y is $Y=N+X$ for timelike N and spacelike X, so that (8.68) implies $Y^\mu Z^\nu G_{\mu\nu}=0$ for arbitrary vectors Y and Z. This is obviously equivalent to $G_{\mu\nu}=0$. □

The simplest, perhaps somewhat trivial illustration of this formalism is Minkowski space $(\mathbb{M},\eta)$, foliated as $M=\sqcup_{t\in\mathbb{R}}\Sigma_t$, where $\Sigma_t=\{(x^0,\vec{x})\mid x^0=t\}$, corresponding to the time function

$$\mathrm{t}(x^0,\vec{x})=x^0. \tag{8.70}$$

In the usual coordinates one has $g=\eta$, so for this foliation the lapse and the shift are simply

$$L=1; \qquad\qquad S=0. \tag{8.71}$$

[354] This is the case, for example, if $T_{\mu\nu}$ is obtained from a matter action S_M via (7.77), where S_M is obtained by ***minimal coupling***, in that in some special relativistic action, $\eta_{\mu\nu}$ and ∂_μ are replaced by $g_{\mu\nu}$ and ∇_μ, respectively. See Anderson (1981) and Read, Brown, & Lehmkuhl (2018) for interesting perspectives on minimal coupling.

[355] The theorem, due to Moncrief & Teitelboim (1973), is valid with and without the words between brackets.

Furthermore, if we take $\Sigma = \Sigma_0$ as our Cauchy surface–which it clearly is–then the induced initial data on Σ are $\tilde{g}_{ij} = \delta_{ij}$ and, since the fd normal $\vec{N} = (1,0,0,0)$ is independent of (x^1,x^2,x^3) (and even of x^0), we have $\tilde{k}_{ij} = 0$. Let us not fail to notice that these initial data satisfy the constraints (7.148) - (7.149), i.e. (8.65) and (8.67) in vacuum ($E = 0$ and $P_i = 0$). From this, we recover the (Minkowski) metric on any other Σ_t by solving (8.61) - (8.62) with (8.71), that is,

$$\partial_t \tilde{k}_{ij} = \tilde{R}_{ij} + \mathrm{Tr}\,(\tilde{k})\tilde{k}_{ij} - 2\tilde{k}_{ij}^2; \tag{8.72}$$

$$\partial_t \tilde{g}_{ij} = -2\tilde{k}_{ij}, \tag{8.73}$$

with initial conditions $\tilde{g}_{ij}(0) = \delta_{ij}$ and $\tilde{k}_{ij}(0) = 0$, and $\tilde{R}_{ij}(t)$ seen as function of $\tilde{g}_{ij}(t)$. The unique solution is $\tilde{g}_{ij}(t) = \delta_{ij}$ and $\tilde{k}_{ij}(t) = 0$ for all $t \in \mathbb{R}$, upon which (8.14) gives $g_{\mu\nu} = \eta_{\mu\nu}$.

Now make this example nontrivial by considering the curious space-time $(I^+(0), \eta)$, i.e.

$$M = I^+(0) \tag{8.74}$$

is the interior of the forward lightcone emanating from the origin in Minkowski space-time, with (relative) Minkowski metric. For ease of visualization we take $d = 2+1$, and set

$$x^0 = t\cosh(\rho); \qquad x^1 = t\sinh(\rho)\cos(\varphi); \qquad x^2 = t\sinh(\rho)\sin(\varphi), \tag{8.75}$$

where $t > 0$, $\rho \in \mathbb{R}$, and $\varphi \in [0, 2\pi)$. Then define $\Sigma_t \subset I^+(0)$ as the hyperboloid (4.88), i.e.

$$\Sigma_t = H^2_t = \{(x^0,x^1,x^2) \in I^+(0) \mid (x^0)^2 - (x^1)^2 - (x^2)^2 = t^2\}, \tag{8.76}$$

so that $M = \sqcup_{t>0}\Sigma_t$. If we take $\Sigma = \Sigma_1$ to be our Cauchy surface in $I^+(0)$, with initial data

$$\tilde{g} = d\rho^2 + \sinh^2(\rho)d\varphi^2; \tag{8.77}$$

$$\tilde{k} = -\tilde{g} = -d\rho^2 - \sinh^2(\rho)d\varphi^2, \tag{8.78}$$

then $(\tilde{g},\tilde{k})$ satisfy the vacuum constraints (7.148) - (7.149). To check this, one may use (4.85) with $n = 2$ and $k = -1$, so that $\tilde{R}_{ij} = -\tilde{g}_{ij}$ and $\tilde{R} = -2$. Secretly the initial data (8.77) - (8.78) were obtained from the Minkowski metric η expressed in the coordinates (t,ρ,φ), which is

$$\eta = -dt^2 + t^2(d\rho^2 + \sinh^2(\rho)d\varphi^2), \tag{8.79}$$

This trivially reproduces $\tilde{g}$ in (8.77), and also leads to $\tilde{k}$ via the fact that the normal of Σ_t is

$$\vec{N} = (\cosh(\rho), \sinh(\rho)\cos(\varphi), \sinh(\rho)\sin(\varphi)). \tag{8.80}$$

which happens to be independent of t. To recover the Minkowski metric from the initial data $(\Sigma_1,\tilde{g},\tilde{k})$, we once again choose (8.71), as will be justified *a posteriori*, and then solve (8.72) - (8.73) subject to these initial data. A nontrivial computation shows that the solution is given by

$$\tilde{g}(t) = t^2(d\rho^2 + \sinh^2(\rho)d\varphi^2); \tag{8.81}$$

$$\tilde{k}(t) = -t^{-1}\tilde{g}_t = -t(d\rho^2 + \sinh^2(\rho)d\varphi^2). \tag{8.82}$$

Once again using (8.14) with (8.71), this duly recovers the space-time metric (8.79).

The cosmological FLRW solution provides another illustration of the $3+1$ formalism. Short of giving the whole story, our starting point is that homogeneity and isotropy imply that the $3d$ Riemannian manifold $(\Sigma, \tilde{g})$ carrying the initial data is one of the three spaces $(\Sigma_C, \tilde{g}_C)$ of constant curvature studied in §4.4. These spaces are parametrized by $C \in \{-1,0,1\}$, i.e.[356]

$$\Sigma_{-1} = H^3; \qquad \Sigma_0 = \mathbb{R}^3; \qquad \Sigma_1 = S^3. \tag{8.83}$$

The associated 4-metric is given by

$$g = -dt^2 + a(t)^2 g_C, \tag{8.84}$$

where the scale factor $t \mapsto a(t)$, initially defined on $\mathbb{R}^+_*$, is to be determined on the basis of the Einstein equations and the specification of some energy-momentum tensor $T_{\mu\nu}$. The latter is assumed to take the perfect fluid form (7.73), where $u^\mu = (1,0,0,0)$, so that

$$T_{\mu\nu} = \mathrm{diag}(\varepsilon, p, p, p). \tag{8.85}$$

Accordingly $E = \varepsilon$, $P_\mu = 0$, and $S_{ij} = p\delta_{ij}$, cf. (7.75). Since $L = 1$ and $S = 0$, it follows from (8.23) and (8.84), where

$$\tilde{g} = a^2 g_C, \tag{8.86}$$

that

$$\tilde{k}_{ij} = -(\dot{a}/a)\tilde{g}_{ij}, \tag{8.87}$$

and hence

$$\mathrm{Tr}\,(\tilde{k}) = -3\dot{a}/a. \tag{8.88}$$

Here a depends only on t (i.e. it is constant on Σ_C), so that

$$\nabla_l \tilde{k}_{ij} = -(\dot{a}/a)\tilde{\nabla}_l \tilde{g}_{ij} = 0, \tag{8.89}$$

as well as

$$\nabla_i \mathrm{Tr}\,(\tilde{k}) = \partial_i \mathrm{Tr}\,(\tilde{k}) = 0. \tag{8.90}$$

Since $P_\mu = 0$, the momentum constraint (8.67) reads $0 = 0$ and hence is satisfied. Noting that

$$\tilde{R} = 6C/a^2 \tag{8.91}$$

from (4.85), the Hamiltonian constraint (7.148) becomes

$$\frac{C}{a^2} + \left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi}{3}\varepsilon. \tag{8.92}$$

Since eq. (8.60) has been incorporated, what remains is (8.59). After some reshuffling, including removing the $\tilde{R}$ term using (8.92), contracting with $\tilde{g}^{ij}$ gives the second Friedman equation

$$\frac{\ddot{a}}{a} = -\frac{4\pi}{3}(\varepsilon + 3p). \tag{8.93}$$

Textbooks show how to solve (8.92) - (8.93), supplemented with an equation of state (such as $p = 0$, describing dust, or $p = \varepsilon/3$ for photons). One needs considerable philosophical skill and courage to deny that the ensuing expansion of the universe is a real process in time! See §8.11.

[356] To avoid confusion with the extrinsic curvature $\tilde{k}_{ij}$ we now write C for the constant (curvature) k in §4.4.

8.4 Static, stationary, and asymptotically flat space-times

From an expanding universe, we now move to the opposite end of the dynamical spectrum. A space-time (M,g) might naively be called ***stationary*** if there is a globally defined complete *timelike* Killing vector field X for g, and ***static*** if in addition, M admits a foliation à la (8.1) for which X is orthogonal to each leaf Σ_t. By the Frobenius theorem this is the case iff

$$dX^\flat \wedge X^\flat = 0; \qquad \Leftrightarrow \qquad X_{[\mu}\nabla_\nu X_{\rho]} = 0, \tag{8.94}$$

since this expresses the property that the distribution of all vector fields orthogonal to X is integrable. This has the following consequences for the metric.[357] First, g is stationary with respect to X iff at least away from the zeros of X (if any), in coordinates where $X = \partial_t$,

$$g = -L^2(dt+\theta)^2 + \tilde{g}; \qquad \theta = \theta_i dx^i; \qquad \tilde{g} = \tilde{g}_{ij}dx^i dx^j, \tag{8.95}$$

where L, θ_i, and $\tilde{g}_{ij}$ are independent of t. The static case then has $\theta = 0$, i.e.

$$g = -L^2 dt^2 + \tilde{g}, \tag{8.96}$$

where coordinates are such that $x = (t,\vec{x})$ lies in Σ_t as in (8.1), and $\vec{x}$ are coordinates on $\Sigma_t \cong \Sigma$.

Thus *a metric is static iff it is stationary and invariant under time inversion* $t \mapsto -t$. The exterior Schwarzschild solution (9.15) for $r > 2m$ is static, whereas the Kerr metric is stationary; time inversion makes the black hole rotate the other way round, both with $X = \partial_t$, But if we extend the Schwarzschild solution to $0 < r < 2m$, as explained in §9.2, then ∂_t becomes lightlike at $r = 2m$ and even spacelike when $0 < r < 2m$, so that the definition of a stationary space-time has to be relaxed if we wish to cover such cases. This is done as follows (see part 4):[358]

Definition 8.4 *1. A 3d* **Riemannian manifold** $(\Sigma,\tilde{g})$ *is called* **asymptotically flat** *if:*

(i) There is a bounded set $K \subset \Sigma$ whose complement $\Sigma\backslash K$ is a finite union of **ends** Σ_α^{ext}*, each of which is diffeomorphic to $\mathbb{R}^3\backslash B_1^3$ (where $B_1^3 = \{\vec{x}\in\mathbb{R}^3 \mid x^2+y^2+z^2 \le 1\}$).*

(ii) For each $\alpha = 1,\dots,\ell$ there exists a coordinate chart $\varphi_\alpha : \Sigma_\alpha^{ext} \xrightarrow{\cong} \mathbb{R}^3\backslash B_1^3$ in which the 3-metric $g^{(\alpha)} = \tilde{g}_{|\Sigma_\alpha^{ext}}$ is asymptotically Euclidean in the sense that, pointwise as $|\vec{x}| \to \infty$,

$$|\tilde{g}_{ij}^{(\alpha)}(\vec{x}) - \delta_{ij}| + |\vec{x}|\,|\partial_k \tilde{g}_{ij}^{(\alpha)}(\vec{x})| + |\vec{x}|^2\,|\partial_k\partial_l \tilde{g}_{ij}^{(\alpha)}(\vec{x})| = O(|\vec{x}|^{-1}). \tag{8.97}$$

(iii) The Ricci scalar $\tilde{R}$ of $\tilde{g}$ is integrable, i.e. $\int_\Sigma \omega_g|\tilde{R}| = \int_\Sigma d^3x\,\sqrt{\det(\tilde{g})}(x)|\tilde{R}(x)| < \infty$.

[357] Here we follow Chruściel (2020), §4.3.1 and then §4.3.7, where omitted details are simple exercises.

[358] Such definitions go back at least to Lichnerowicz (1955) and have been made increasingly precise afterwards. For part 1 see Lee (2019), Definition 3.5, in which we take the simplest decay conditions. One may generalize $O(|\vec{x}|)^{-1}$ in (8.97) to $O(|\vec{x}|^{-p})$ for some $p \in (\frac{1}{2},1]$, in which case (8.98) generalizes $O(|\vec{x}|^{-2})$ to $O(|\vec{x}|^{-p-1})$; one needs $p > 1/2$ for the asymptotic mass Π^0 in (8.103) to exist, and $p \le 1$ for it to be potentially nonzero. In the presence of matter one furthermore requires $|E(\vec{x})| + |P_i(\vec{x})| = O(|\vec{x}|^{-3})$, cf (8.65) and (8.67), cf. Cederbaum & Sakovich (2018). But if the constraints (8.65) - (8.67) and the dominant energy condition hold, this may be replaced by our condition (iii), taken from Schoen (2009), Lecture 9, which condition is very convenient in practice. A completely different way of defining asymptotic flatness, going back to Penrose, will be discussed in §10.3.

2. *An* **initial data set** $(\Sigma,\tilde{g},\tilde{k})$ *is* **asymptotically flat** *if* $(\Sigma,\tilde{g})$ *is, and* $\tilde{k}^{(\alpha)}=\tilde{k}_{|\Sigma^{ext}_{\alpha}}$ *satisfies*

$$|\tilde{k}^{(\alpha)}_{ij}(\vec{x})|+|\vec{x}|\,|\partial_k\tilde{k}^{(\alpha)}_{ij}(\vec{x})|=O(|\vec{x}|^{-2}). \tag{8.98}$$

3. *A* **space-time** (M,g) *is* **asymptotically flat** *if it has a spacelike hypersurface* $\imath:\Sigma\hookrightarrow M$ *for which the induced data set* $(\Sigma,\tilde{g},\tilde{k})$ *given by the induced 3-metric* $\tilde{g}=\imath^*g$ *and the second fundamental form* $\tilde{k}$ *of the embedding* $\imath$ *is asymptotically flat (as in items 1–2).*

4. *An asymptotically flat space-time is* **stationary** *if it has a complete Killing vector field* X *that at each end* Σ^{ext}_{α} *is timelike, and* L *and* θ_i *in* (8.95) *are* $O(|\vec{x}|^{-1})$ *as in* (8.97)*. It is* **static** *if* X *in addition satisfies the integrability condition* (8.94)*, so that* $\theta=0$*.*

The idea is that the Killing vector field X defining stationarity need only be timelike "far away". In asymptotically flat stationary space-times, the flow φ_t of X (assumed complete) consist of isometries and since far-away observers with four-velocity $u=X/\sqrt{-g(X,X)}$ (who consider themselves at rest) move along the flow lines, they see no change. One does need the full complexity of this definition, since already the maximally extended Schwarzschild space-time (i.e. the Kruskal solution) has two ends. Noting that in coordinates where (8.96) holds the shift S vanishes, it should be clear from (8.61) - (8.62) that the static case, simply corresponds to

$$\tilde{k}_{ij}=0. \tag{8.99}$$

In that case at least in *vacuo* the momentum constraint (8.66) is identically satisfied, whereas the dynamical Einstein equation (8.61) and the Hamiltonian constraint (8.65) simplify to

$$\tilde{\nabla}_i\tilde{\nabla}_jL=\tilde{R}_{ij}L; \tag{8.100}$$

$$\tilde{R}=0, \tag{8.101}$$

respectively. Contracting (8.100) with $\tilde{g}^{ij}$ and using (8.101) gives $\tilde{\Delta}_{\tilde{g}}L=0$, where $\tilde{\Delta}_{\tilde{g}}:=\tilde{g}^{ij}\tilde{\nabla}_i\tilde{\nabla}_j$ is the $3d$ Laplacian determined by $\tilde{g}$. In the presence of (8.100), this is equivalent to (8.101), so that the Einstein equations for a static space-time are also given by

$$\tilde{\nabla}_i\tilde{\nabla}_jL=\tilde{R}_{ij}L; \qquad \tilde{\Delta}_{\tilde{g}}L=0. \tag{8.102}$$

The oldest rigorous result in this context is ***Lichnerowicz's theorem*** from 1939, which states that if (M,g) is static, asymptotically flat, *and geodesically complete* (a property the Schwarzschild space-time lacks), then $(\Sigma,\tilde{g})$ is isometric to flat Euclidean space and $L=1$, so that (M,g) is isometric to Minkowski space-time.[359] This follows from the theory of the Laplace equation and the boundary condition $L\to 1$ at spatial infinity, cf. (8.96) and (8.97).

More generally, any *geodesically complete* stationary space-time solving the vacuum Einstein equations is isometric to $\mathbb{R}\times\Sigma$ with flat metric,[360] so that the assumption of asymptotic flatness in Lichnerowicz's theorem is only needed to enforce $\Sigma\cong\mathbb{R}^3$. See also Theorem 10.25 in §10.9.

As will be justified below from physics,[361] the ***asymptotic (ADM) energy*** Π^0 is defined by

$$\Pi^0:=\frac{1}{16\pi}\lim_{r\to\infty}\int_{S^2_r}d^2\sigma^i\,(\partial_j\tilde{g}_{ij}-\partial_i\tilde{g}_{jj}), \tag{8.103}$$

where $d^2\sigma^i=x^i\sin\theta d\theta d\varphi$, with $\vec{x}=(r\sin\theta\cos\varphi,r\sin\theta\sin\varphi,r\cos\theta)$ as usual.

[359] Choquet-Bruhat (2018) reports that this result even impressed Einstein, who had been unable to prove it.
[360] See Anderson (2000a), Chruściel, Lopes Costa, & Heusler (2012), and Cortier & Minerbe (2016).
[361] There are *many* other concepts of "mass" in GR, reviewed by Galloway, Miao, & Schoen (2015) and Lee (2019).

The most famous "elliptic PDE" result in mathematical GR is the ***positive mass theorem***:[362]

Theorem 8.5 *Let $(\Sigma,\tilde{g})$ be (geodesically) complete and asymptotically flat, with $\tilde{R}\geq 0$. Then $\Pi^0\geq 0$, with equality $\Pi^0=0$ iff $(\Sigma,\tilde{g})$ is isometric to Euclidean space $(\mathbb{R}^3,\delta)$.*

The assumption $\tilde{R}\geq 0$ may be motivated by noting that in static space-times one has

$$\tilde{R}=16\pi E, \tag{8.104}$$

which is the Hamiltonian constraint (8.65) with $\tilde{k}=0$. Compare with the Newtonian formula

$$\Delta V=4\pi\rho \tag{8.105}$$

for the gravitational potential, with the difference that (8.105) *determines* V, whereas (8.104) merely *constrains* the metric $\tilde{g}_{ij}$. In any case, $\tilde{R}\geq 0$ now simply comes down to $E\geq 0$. In this light, one may also justify (8.103) by noting that for asymptotically flat spaces one has

$$\tilde{R}=\partial_i(\partial_j\tilde{g}_{ij}-\partial_i\tilde{g}_{jj})+O(|x|^{-4}), \tag{8.106}$$

where the first term comes from the first two terms in (7.26), in $d=3$ of course, and the last comes from the $\Gamma\cdot\Gamma$ terms, where Γ contains first derivatives of $\tilde{g}$ and hence is $O(|x|^{-2})$. Assuming for the moment that (8.106) holds globally and that $\Sigma=\mathbb{R}^3$, the total energy Π^0 of all matter plus the gravitational field may then be defined as

$$\Pi^0:=\lim_{r\to\infty}\int_{B^3_r}d^3xE=\frac{1}{16\pi}\lim_{r\to\infty}\int_{B^3_r}d^3x\tilde{R}=\frac{1}{16\pi}\lim_{r\to\infty}\int_{S^2_r}d^2\sigma^i\,(\partial_j\tilde{g}_{ij}-\partial_i\tilde{g}_{jj}), \tag{8.107}$$

which recovers (8.103).[363] For example, for the spatial part of the Schwarzschild metric, i.e.

$$\tilde{g}=\left(1-\frac{2m}{r}\right)^{-1}dr^2+r^2(d\theta^2+\sin^2\theta d\varphi^2), \tag{8.108}$$

one obtains $\Pi^0=m$.[364] Similarly, the ***asymptotic (ADM) momentum*** is defined by

$$\Pi_j:=\frac{1}{8\pi}\lim_{r\to\infty}\int_{S^2_r}d^2\sigma^i\,\tilde{\pi}_{ij}, \tag{8.109}$$

where the canonical momentum $\tilde{\pi}_{ij}$ is defined in terms of $\tilde{g}_{ij}$ and $\tilde{k}_{ij}$ by (8.209) below in §8.7. This leads to a generalization of Theorem 8.5. Let asymptotically flat initial data $(\Sigma,\tilde{g},\tilde{k})$ satisfy

$$\tfrac{1}{2}(\tilde{R}-\mathrm{Tr}\,(\tilde{k}^2)+\mathrm{Tr}\,(\tilde{k})^2)\geq\|\tilde{\nabla}_j\tilde{k}^j_i-\tilde{\nabla}_i\mathrm{Tr}\,(\tilde{k})\|_{\tilde{g}}. \tag{8.110}$$

Then $\Pi^0\geq\|\vec{\Pi}\|$. If the constraints (8.65) - (8.67) hold, eq. (8.110) is equivalent to $E\geq\|\vec{P}\|$.

[362] The original proof is due to Schoen & Yau (1979, 1981); see also Schoen (1989, 2009). For spin manifolds Witten (1981) and Parker & Taubes (1982) proved the theorem in a completely different way. See also Lee (2019) for both proofs. The ***Riemannian Penrose inequality*** sharpens the positive mass theorem; see §10.11.

[363] Integrability of $\tilde{R}$ and existence of Π^0 are equivalent, and since the former is in Definition 8.4, the latter exist.

[364] See Poisson (2004), §4.3.2, Gourgoulhon (2012), §8.3, Example 8.1, or Schoen (2009), Lecture 9. An efficient way to do this computation, following the latter, is to write $\tilde{g}_{ij}=(1+m/2|x|)^4\delta_{ij}+O(1/|x|^2)$. This gives the integrand as $x^i(\partial_j\tilde{g}_{ij}-\partial_i\tilde{g}_{jj})=4m(1+m/2|x|)^3/|x|+O(1/|x|^2)$. As $r\to\infty$ the error term does not contribute, whilst the first gives $\int_{S^2_r}d^2\sigma^i\,(\partial_j\tilde{g}_{ij}-\partial_i\tilde{g}_{jj})=16\pi m(1+m/2r)^3$, which as $r\to\infty$ yields $16\pi m$, and hence $\Pi^0=m$.

The proof of Theorem 8.5 (which implies the generalization) is lengthy and difficult, but the main steps are as follows. First, as explained in more detail in §8.6 one may apply a conformal transformation $\hat{g} = \Omega^4 \tilde{g}$, where the strictly positive function $\Omega \in C^\infty(\Sigma)$ solves the linear PDE

$$(\tilde{\Delta} - \tfrac{1}{8}\tilde{R})\Omega = 0. \tag{8.111}$$

Then $\hat{R} = 0$, where $\hat{R}$ is the Ricci scalar for $\hat{g}$. Since the flat space Laplace equation $\Delta f = 0$ in $d = 3$ has fundamental solution $f = C/r$, we obtain $\Omega = 1 + C/r + O(1/r^2)$. The point is that $C < 0$, as follows by integrating the equality $\tilde{\Delta}\,\Omega = \frac{1}{8}\tilde{R}\Omega$ over a three-ball B^3 and using $\tilde{R} \geq 0$. Hence $\Pi^0(\hat{g}) = \Pi^0(\tilde{g}) - C \leq \Pi^0(\tilde{g})$, which reduces the proof to the case $\tilde{R} = 0$.

The proof then proceeds by contradiction. If $\Pi^0 < 0$, then one can find a smooth asymptotically flat metric $\breve{g}$ that equals $\tilde{g}$ in B^3_ρ for some $\rho > 0$ and equals $\breve{\Omega}^4\delta$ outside B^3_ρ; this works because, as before, $\breve{\Omega} = 1 + C'/r + O(1/r^2)$ for some $C' < 0$, and we now have $C = \Pi^0$. This metric can, in turn, be used to construct a new metric g' that is *exactly* Euclidean outside some three-ball and has $R' > 0$. This, however, contradicts the following remarkable lemma:[365]

Lemma 8.6 *If a Riemannian manifold with Ricci scalar $R \geq 0$ is isometric to Euclidean space $\mathbb{R}^n \backslash B^n_\rho$ outside some compact set (for some $\rho > 0$), then it is isometric to $(\mathbb{R}^n, \delta)$.*

This argument proves the first part of Theorem 8.5. A very elegant argument for the second claim comes from ***Ricci Flow***, the technique used to prove the Poincaré conjecture.[366] Here (especially in $d = 3$) a "time" dependent Riemannian metric $\tilde{g}(t)$ satisfies the parabolic PDE

$$\frac{\partial \tilde{g}_{ij}}{\partial t} = -2\tilde{R}_{ij}, \tag{8.112}$$

from some given initial metric $\tilde{g}(0)$. This induces a flow of the Ricci scalar $\tilde{R}$, namely

$$\partial_t \tilde{R} = \tilde{\Delta}\,\tilde{R} + 2\tilde{R}_{ij}\tilde{R}^{ij}. \tag{8.113}$$

It can then be shown that Π^0 is independent of t (which is not surprising since it is an asymptotic quantity), so if $\Pi^0 = 0$ for $g(0)$, then $\Pi^0 = 0$ for all $g(t)$. Step 1 above then shows that $\tilde{R}(t) = 0$ and hence (8.113) yields $R_{ij} = 0$. In $d = 3$ this means that the Riemann tensor also vanishes (see §4.5) and hence by Theorem 4.1 our space is locally Euclidean. Asymptotic flatness then prevents nontrivial topology for large r, whereas geodesic completeness forces the bounded set $K \subset \Sigma$ in Definition 8.4.1 to be compact. Lemma 8.6 finally yields the claim. □

Towards a further (covariant) justification of the definitions (8.103) and (8.109), in the physics literature on linearized gravity, asymptotic flatness is expressed by the decomposition

$$g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu}, \tag{8.114}$$

where $\eta_{\mu\nu}$ is the Minkowski metric and $h_{\mu\nu}$ is "small"; this *Ansatz* seems predicated on the topological assumption $M \cong \mathbb{R}^4$. Relating this to the assumptions in Definition 8.4 is highly nontrivial,[367] but since we only try to *motivate* (8.103) and (8.109) we omit the details.

[365] See Corollary 2.32 in Lee (2019). This result, first conjectured by Geroch in 1975, is equivalent to the nonexistence of positive scalar curvature metrics on a torus, which is Theorem 1.30 in Lee (2019).

[366] The Poincaré conjecture states that any compact simply connected 3-manifold is diffeomorphic to the three-sphere S^3. It was proved by the eccentric Russian mathematician Perelman in 2002–2003. See Morgan & Tian (2007). The Ricci Flow approach to the positive mass theorem was developed by McFeron & Székelyhidi (2012).

[367] See Christodoulou & Klainerman (1993) and Klainerman & Nicolò (2003a). See also Weinberg (1972), §7.6, Misner, Thorne, & Wheeler (1973), §20.2, Jaramillo & Gourgoulhon (2009), and de Haro (2021).

Assuming (8.114) and topological triviality of M as above, we can expand the Einstein tensor $G_{\mu\nu}$ to linear order in h, calling the result $G^L_{\mu\nu}$. In terms of the convenient expression

$$\bar{h}^{\mu\nu} := h^{\mu\nu} - \tfrac{1}{2}\eta^{\mu\nu}\eta_{\rho\sigma}h^{\rho\sigma}, \tag{8.115}$$

where indices are raised and lowered through the Minkowski metric η, this gives

$$G^L_{\mu\nu} = -\tfrac{1}{2}\Box_\eta \bar{h}_{\mu\nu} - \eta_{\mu\nu}\partial_\alpha\partial_\beta\bar{h}^{\alpha\beta} + \partial_\alpha\partial_\mu\bar{h}^\alpha_\nu + \partial_\alpha\partial_\nu\bar{h}^\alpha_\mu, \tag{8.116}$$

which leads to the linearized Einstein equations

$$G^L_{\mu\nu} \approx 8\pi T_{\mu\nu}. \tag{8.117}$$

For later use, we note that $G^{\mu\nu}_L := \eta^{\mu\rho}\eta^{\nu\sigma}G^L_{\rho\sigma}$ may be written as

$$G^{\mu\nu}_L = \tfrac{1}{2}\partial_\alpha\partial_\beta H^{\mu\alpha\nu\beta}; \tag{8.118}$$

$$H^{\mu\alpha\nu\beta} := \bar{h}^{\alpha\nu}\eta^{\mu\beta} + \bar{h}^{\mu\beta}\eta^{\alpha\nu} - \bar{h}^{\mu\nu}\eta^{\alpha\beta} - \bar{h}^{\alpha\beta}\eta^{\mu\nu}, \tag{8.119}$$

where the so-called "superpotential" H has (anti) symmetries

$$H^{\mu\alpha\nu\beta} = -H^{\alpha\mu\nu\beta} = H^{\mu\alpha\beta\nu}. \tag{8.120}$$

One may also rewrite the full Einstein equations in *exact* form as

$$G^L_{\mu\nu} = 8\pi\tau_{\mu\nu} := 8\pi(T_{\mu\nu} + t_{\mu\nu}); \tag{8.121}$$

$$t_{\mu\nu} := (8\pi)^{-1}(G^L_{\mu\nu} - G_{\mu\nu}), \tag{8.122}$$

where $t_{\mu\nu}$, sometimes seen as the ***self-energy-momentum (pseudo) tensor of the gravitational field***, is quadratic in h. Eqs. (8.118) and (8.120) then immediately give the conservation laws

$$\partial_\mu G^{\mu\nu}_L = 0; \qquad\qquad \partial_\mu\tau^{\mu\nu} = 0, \tag{8.123}$$

where the first one is an identity and the second one relies on the field equation (8.121).

Assuming for the moment that (8.114) holds globally and that $\Sigma = \mathbb{R}^3$, the total energy-momentum Π^μ of all matter plus the gravitational field may then be defined as

$$\Pi^\mu := \lim_{r\to\infty}\int_{B^3_r} d^3x\,\tau^{0\mu} = \frac{1}{8\pi}\lim_{r\to\infty}\int_{B^3_r} d^3x\,G^{0\mu}_L, \tag{8.124}$$

where have used the *exact* equation (8.121). The same expression on the right-hand side appears if we define Π^μ as the integral of $T^{0\mu}$ instead of $\tau^{0\mu}$, and then use the *approximate* (linearized) equations (8.117). So either way, we may proceed using (8.118) and (8.120) to obtain

$$\begin{aligned}\Pi^\mu &= \frac{1}{16\pi}\lim_r\int_{B^3_r} d^3x\,\partial_\alpha\partial_\beta H^{0\alpha\mu\beta} = \frac{1}{16\pi}\lim_r\int_{B^3_r} d^3x\,\partial_i\partial_\beta H^{0i\mu\beta}\\ &= \frac{1}{16\pi}\lim_r\int_{S^2_r} d^2\sigma_i\,\partial_\beta H^{0i\mu\beta},\end{aligned} \tag{8.125}$$

which is valid whatever is going on inside the compact region $K\subset\Sigma$ about which we have no information, so that we may also take (8.125) as the *definition* of Π^μ. In particular, for $\mu = 0$,

$$\begin{aligned}\Pi^0 &= \frac{1}{16\pi}\lim_r\int_{S^2_r} d^2\sigma_i\,\partial_j H^{0i0j} = \frac{1}{16\pi}\lim_r\int_{S^2_r} d^2\sigma^i\,(\partial_j h_{ij} - \partial_i h_{jj})\\ &= \frac{1}{16\pi}\lim_r\int_{S^2_r} d^2\sigma^i\,(\partial_j\tilde{g}_{ij} - \partial_i\tilde{g}_{jj}) = \frac{1}{16\pi}\lim_r\int_{S^2_r}(\nabla_\eta\cdot\tilde{g} - d(\mathrm{Tr}_\eta(\tilde{g}))),\end{aligned} \tag{8.126}$$

which is (8.103). The derivation of (8.109) from (8.124) is similar and is left to the reader.

8.5 The origin of diffeomorphism invariance?

Having seen the linearized Einstein equations (8.117), it would be a pity not to mention an argument for (at least infinitesimal) general covariance that at least sheds new light on this property compared to the kind of arguments mentioned in §1.10. General relativists do not like this argument, since it takes place in Minkowski space-time and puts special relativity before general relativity. But special relativists (i.e. particle physicists) do, for the very same reason![368]

The argument is based on the unitary irreducible representations of the Poincaré group P, which were classified by Wigner (1939).[369] We first define P as the semidirect product

$$P := O(3,1) \ltimes \mathbb{R}^4; \tag{8.127}$$

$$O(3,1) := \{\Lambda \in GL_4(\mathbb{R}) \mid \langle \Lambda x, \Lambda y\rangle_M = \langle x,y\rangle_M \, \forall x,y \in \mathbb{R}^4\}, \tag{8.128}$$

where $\langle \cdot,\cdot,\rangle_M$ is the Minkowski inner product in $\mathbb{R}^4$. One calls $O(3,1)$ the ***Lorentz group***. This means that P consists of pairs $(\Lambda, a) \in O(3,1) \times \mathbb{R}^4$, equipped with group operations

$$(\Lambda,a)\cdot(\Lambda',a') = (\Lambda\Lambda', a + \Lambda\cdot a'); \tag{8.129}$$

$$(\Lambda,a)^{-1} = (\Lambda^{-1}, -\Lambda^{-1}\cdot a). \tag{8.130}$$

The full Lorentz group $O(3,1)$ has four connected components, which may be identified by the (independent) conditions $\det(\Lambda) = \pm 1$ and $\pm\Lambda^0{}_0 > 0$. For the moment we restrict ourselves to the connected component containing the identity (which is like $SO(3)$ in $O(3)$), in which $\det(\Lambda) = 1$ and $\Lambda^0{}_0 > 0$. This group, which we call L, is the ***proper orthochronous Lorentz group***. It gives rise to $P_0 = L \ltimes \mathbb{R}^4$, which is the connected component of the identity in P. If we write the L-action on $\mathbb{R}^4$ as $x^\mu \mapsto \Lambda^\mu{}_\nu x^\nu$, then $\Lambda \in GL(4,\mathbb{R})$ should satisfy

$$\eta_{\alpha\beta}\Lambda^\alpha{}_\mu \Lambda^\beta{}_\nu = \eta_{\mu\nu}. \tag{8.131}$$

Wigner showed that it is the *dual action* of L on $(\mathbb{R}^4)^*$, seen as 4-momentum space, that counts for the classification: if we denote elements of the dual vector space $(\mathbb{R}^4)^* \cong \mathbb{R}^4$ by p_μ, then the dual action is $p_\mu \mapsto \Lambda_\mu{}^\nu p_\nu$, where, as the notation indicates, $\Lambda_\mu{}^\nu = \eta_{\mu\alpha}\eta^{\nu\beta}\Lambda^\alpha{}_\beta$.

Writing $p^2 = -p_0^2 + p_1^2 + p_2^2 + p_3^2$, the L-orbits $\mathscr{O}$ in $(\mathbb{R}^4)^* = \mathbb{R}^4$ are easily seen to be:

1. $\mathscr{O}_0^0 = \{(0,0,0,0)\}$, with stabilizer $L_0 = L$;
2. $\mathscr{O}_m^\pm = \{p \in \mathbb{R}^4 \mid p^2 = -m^2, \pm p^0 > 0\}$, $m > 0$, with stabilizer $L_0 \cong SO(3)$;
3. $\mathscr{O}_0^\pm = \{p \in \mathbb{R}^4 \mid p^2 = 0, \pm p^0 > 0\}$, with stabilizer $L_0 \cong E(2) = SO(2) \ltimes \mathbb{R}^2$;
4. $\mathscr{O}_{im} = \{p \in \mathbb{R}^4 \mid p^2 = m^2\}$, $m > 0$, with stabilizer $L_0 \cong SO(2,1)$.

Here the stabilizers L_0 are found by taking reference points $(\pm m,0,0,0)$ in case 2, $(\pm 1,0,0,-1)$ in case 3, and $(0,0,0,m)$ in case 4. The physically relevant cases seem to be $\mathscr{O}_m^+$ and $\mathscr{O}_0^+$, since $\mathscr{O}_{im}$ describes tachyons, which probably do not exist. The unitary irreducible representations of P_0 are then labeled by pairs $(\mathscr{O},\chi)$, where $\mathscr{O} \subset (\mathbb{R}^4)^*$ is one of these orbits, and χ labels a unitary irreducible representation of the corresponding stabilizer (which depends on $\mathscr{O}$).

[368] See Weinberg (1972), §10.2, Scharf (2016), and, reluctantly, Misner, Thorne, & Wheeler (1973), Box 17.2 (5).
[369] See also Barut & Raçka (1977), chapter 17, and Landsman (1998), §IV.3.

In particular, for $\mathscr{O}_m^+$ with $m>0$ one has $\chi = j \in \{0,1,2,\ldots\}$. This describes the *spin* of the (elementary) particle described by the given representation,[370] so that the total label is (m,j). For $m=0$, on the other hand, χ labels unitary irreducible representations of the $2d$ Euclidean group $E(2)$. This would in principle involve another continuous label, but it seems that in reality only the case occurs where the elements of $\mathbb{R}^2$ are represented trivially, so that one just needs a label for $SO(2)$, which is $\lambda \in \mathbb{Z}$, called *helicity*. Denoting elements of $E(2) = SO(2) \ltimes \mathbb{R}^2$ by (z,x,y), where $z \in SO(2) \cong \mathbb{T}$ and $(x,y) \in \mathbb{R}^2$, helicity is just the character

$$u_\lambda : E(2) \to \mathbb{T}; \qquad u_\lambda(z,x,y) = z^\lambda. \tag{8.132}$$

Including parity, i.e. $\mathrm{diag}(1,-1,-1,-1)$, in L then forces λ to be accompanied by $-\lambda$. The case $\lambda = 0$ does not occur, but the pair $\lambda = \pm 1$ describes ***photons*** whereas $\lambda = \pm 2$ gives ***gravitons***.

In this light, traditional (relativistic) quantum field theory may be understood as follows:

1. *Linear* field equations distill specific unitary irreducible representations (typically realized in momentum space) from covariantly transforming fields (defined in space-time);

2. Nonlinear terms (often dictated by other symmetries than Poincaré invariance) are added to the equations to describe interactions between the elementary particles thus involved.[371]

For example, the Klein–Gordon equation $(\Box - m^2)\varphi = 0$, where $\Box = -\partial_t^2 + \Delta$ and $\varphi : \mathbb{R}^4 \to \mathbb{R}$ is a real scalar field, selects the representation labeled by $(m,+,j=0)$. Namely, we write

$$\varphi(x) = \int_{\mathbb{R}^3} \frac{d^3\mathbf{p}}{(2\pi)^3 p^0}\, e^{ipx}\hat{\varphi}(\mathbf{p}), \tag{8.133}$$

where $px = p_\mu x^\mu$ with $p^0 = \sqrt{|\mathbf{p}|^2 + m^2}$, which solves the Klein–Gordon equation, and take

$$\hat{\varphi} \in H = L^2(\mathbb{R}^3, d^3\mathbf{p}/p^0), \tag{8.134}$$

whose measure d^3p/p^0 is Lorentz-invariant. The natural space-time covariant P-action

$$(\Lambda,a)\cdot\varphi(x) = \varphi(\Lambda^{-1}(x-a)) \tag{8.135}$$

then corresponds to Wigner's realization of the $(\mathscr{O}_m^+,0)$ unitary irreducible representation, i.e.

$$U_{(m,+,0)}(\Lambda,a)\hat{\varphi}(\mathbf{p}) = e^{-ipa}\hat{\varphi}(\Lambda^{-1}\mathbf{p}). \tag{8.136}$$

For $m>0$ this can be generalized to arbitrary spin $j>0$, but our interest lies in the case $m=0$, where we again recall Wigner's manifestly unitary (but unrecognizably space-time covariant) formula for the representation $U_{(0,+,\lambda)}(P)$ labeled by the orbit $\mathscr{O}_0^+$ and the helicity $\lambda \in \mathbb{Z}$. For any λ this is realized on the same Hilbert space (8.134), but now (as in the case of higher spin) Wigner's formula for the explicit realization requires a (measurable) cross-section

$$b : \mathscr{O}_0^+ \to L \tag{8.137}$$

[370] By another analysis of Wigner, one should also include representations of the universal covering group of P, which allows j to be half-integral, too. Likewise for helicity, but this complication is not needed in what follows.

[371] The reason general relativists frown on this is precisely that in GR it is the full nonlinear (Einstein) field equations that are more natural than the linear ones, and a similar point in fact applies to Yang–Mills theories.

of the canonical projection

$$\pi : L \to \mathscr{O}_0^+ \cong L/E(2), \tag{8.138}$$

i.e. $\pi \circ b = \mathrm{id}_{\mathscr{O}_0^+}$, where without loss of generality we may and will assume that

$$b(p') = e, \tag{8.139}$$

the unit element of L. The (L-equivariant) diffeomorphism $\mathscr{O}_0^+ \cong L/E(2)$ is $p \mapsto [\Lambda]$, where $\Lambda \in L_0$ satisfies $p = \Lambda p'$ with $p'_\mu = (1,0,0,1)$ and $[\Lambda]$ is its equivalence class (i.e. image under π) in $L/E(2)$. Moreover, we have $\mathscr{O}_0^+ \cong \mathbb{R}^3$ by mapping $(p^0, \mathbf{p}) \in \mathscr{O}_0^+$ with $p^0 = |\mathbf{p}|$ to $\mathbf{p} \in \mathbb{R}^3$; this diffeomorphism is also L-equivariant if we define $\Lambda\mathbf{p}$ as the spatial part of $\Lambda_\mu{}^\nu p_\nu$. One then verifies that the ***Wigner cocycle*** $b(\mathbf{p})\Lambda b(\Lambda^{-1}\mathbf{p})$ lies in $L_0 = E(2)$. Then for any $(\Lambda, a) \in P_0$,

$$U_{(0,+,\lambda)}(\Lambda, a)\hat{\psi}(\mathbf{p}) = e^{-ipa} u_\lambda(b(\mathbf{p})\Lambda b(\Lambda^{-1}\mathbf{p}))\hat{\psi}(\Lambda^{-1}\mathbf{p}). \tag{8.140}$$

For $\lambda = \pm 1$, we now relate this unitary yet mysterious expression to the manifestly space-time covariant action of P_0 on the electromagnetic field potential A_μ, on which we simply have

$$(\Lambda, a) \cdot A_\mu(x) = \Lambda_\mu{}^\nu A_\nu(\Lambda^{-1}(x-a)). \tag{8.141}$$

The idea is that we pass to a new space, consisting of *solutions of the Maxwell equation*

$$\Box A_\mu - \partial_\mu(\partial^\nu A_\nu) = 0, \tag{8.142}$$

cf. (7.87)*, modulo gauge transformations*

$$A_\mu \mapsto A_\mu + \partial_\mu \lambda. \tag{8.143}$$

Both the equations and the quotienting are Poincaré invariant: P maps the solution space to (8.142) to itself, and if $A_\mu \sim A'_\mu$ in that $A_\mu = A'_\mu + \partial_\mu\lambda$, then $(\Lambda, a) \cdot A_\mu \sim (\Lambda, a) \cdot A'_\mu$ for any $(\Lambda, a) \in P$. The second (quotienting) step may, in turn, be performed in two stages:

1. Find a representative A_μ of A'_μ in its equivalence class under (8.143) by imposing the Lorenz gauge $\partial^\nu A_\nu = 0$, cf. (7.92). This can be done by solving λ from $\Box\lambda = -\partial^\nu A'_\nu$.

2. Quotient by the residual gauge transformations within some class of solutions of the pair

$$\Box A_\mu = 0; \qquad \partial^\nu A_\nu = 0. \tag{8.144}$$

The λ in the residual gauge transformations (8.143) following (8.144) should then satisfy

$$\Box\lambda = 0. \tag{8.145}$$

The first equation in (8.144) is solved by the *spatial* Fourier expansion

$$A_\mu(x) = \int_{\mathbb{R}^3} \frac{d^3\mathbf{p}}{(2\pi)^3 p^0} e^{ipx} \hat{A}_\mu(\mathbf{p}), \tag{8.146}$$

where this time $p^0 = |\mathbf{p}|$ and each component $\hat{A}_\mu \in H$ as in (8.134). The second equation in (8.144) comes down to $p^\mu \hat{A}_\mu(\mathbf{p}) = 0$. Under Lorentz transformations, from (8.141) we have

$$(\Lambda, a) \cdot \hat{A}_\mu)(\mathbf{p}) = e^{-ipa} \Lambda_\mu{}^\nu \hat{A}_\nu(\Lambda^{-1}\mathbf{p}). \tag{8.147}$$

To make this look more like Wigner's unitary expression (8.140) we change variables to

$$\tilde{A}_\mu(\mathbf{p}) = (b(\mathbf{p})^{-1})_\mu{}^\nu \hat{A}_\nu(\mathbf{p}), \tag{8.148}$$

so that (8.147) becomes

$$(\Lambda, a)\cdot\tilde{A}_\mu(\mathbf{p}) = e^{-ipa}(b(\mathbf{p})\Lambda b(\Lambda^{-1}\mathbf{p}))_\mu{}^\nu \tilde{A}_\nu(\Lambda^{-1}\mathbf{p}). \tag{8.149}$$

Taking $\mathbf{p} = (0,0,1) \equiv \mathbf{p}'$ and hence $(p^0, \mathbf{p}) = p'$, and $\Lambda \in E(2)$, eqs. (8.140) and (8.149) become

$$U_{(0,+,\lambda)}(\Lambda, 0)\hat{\psi}(\mathbf{p}') = u_\lambda(\Lambda)\hat{\psi}(\mathbf{p}'); \tag{8.150}$$

$$(\Lambda, 0)\cdot\hat{A}_\mu(\mathbf{p}') = \Lambda_\mu{}^\nu \hat{A}_\nu(\mathbf{p}'), \tag{8.151}$$

where we used (8.139), the (defining) property $\Lambda p' = p'$ for $\Lambda \in L_0 = E(2)$, eq. (8.147), and $\tilde{A}_\mu(\mathbf{p}') = \hat{A}_\mu(\mathbf{p}')$, which follows from (8.148). Therefore, in order to compare the covariant transformation (8.141) with the unitary representation (8.140) all we need to do is look at the solutions $\hat{A}_\mu(\mathbf{p})$ of the two equations (8.144), quotiented by (8.143) at the special point $\mathbf{p} = \mathbf{p}'$. At this point,[372] the second equation in (8.144) and gauge transformation (8.143) become

$$\hat{A}_0(\mathbf{p}') = \hat{A}_3(\mathbf{p}'); \tag{8.152}$$

$$\hat{A}_0(\mathbf{p}') \mapsto \hat{A}_0(\mathbf{p}') + i\hat{\lambda}(\mathbf{p}'); \qquad \hat{A}_3(\mathbf{p}') \mapsto \hat{A}_3(\mathbf{p}') + i\hat{\lambda}(\mathbf{p}'); \tag{8.153}$$

$$\hat{A}_1(\mathbf{p}') \mapsto \hat{A}_1(\mathbf{p}'); \qquad \hat{A}_2(\mathbf{p}') \mapsto \hat{A}_2(\mathbf{p}'), \tag{8.154}$$

where $\hat{\lambda} \in H$, as in (8.134), defines the residual gauge function λ, required after all to satisfy (8.145), by a formula analogous to (8.146). Since $\hat{A}_0$ can be eliminated in favour of $\hat{A}_3$ and the latter is pure gauge, is clear that, frozen at $\mathbf{p}'$, the true unconstrained degrees of freedom after solving (8.144) and quotienting by the (residual) gauge freedom, are $\hat{A}_1$ and $\hat{A}_2$.

The computation of the right-hand side of (8.151) and its comparison with (8.150) relies on the precise embedding of $E(2)$ in L. To describe this, we use a specific basis of the Lie algebra $\mathfrak{l}$, of L, which consists of all real 4×4 matrices M that exponentiate to L; this comes down to the condition $M_{\mu\nu} = -M_{\nu\mu}$, where $M_{\mu\nu} = M_\mu{}^\alpha \eta_{\alpha\nu}$. We then take the following basis of $\mathfrak{l}$:[373]

$$B_1 = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}; \quad B_2 = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}; \quad B_3 = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}; \tag{8.155}$$

$$J_1 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \end{pmatrix}; \quad J_2 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 \end{pmatrix}; \quad J_3 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \tag{8.156}$$

which satisfy the commutation relations (i.e. Lie brackets) appropriate to the Lie algebra $\mathfrak{l}$, viz.

$$[B_i, B_j] = -\varepsilon_{ijk}J_k, \qquad [J_i, J_j] = \varepsilon_{ijk}J_k, \qquad [J_i, B_j] = \varepsilon_{ijk}B_k. \tag{8.157}$$

[372] Of course, one should handle this carefully, since functions in L^2 do not have a value at any particular point. The functional analysis of this situation (and the next) is correctly handled in Landsman & Wiedemann (1995).

[373] Note that these matrices are the $M_\mu{}^\nu = \eta^{\alpha\nu}M_{\mu\nu}$, *not* the $M_{\mu\nu}$ on which the condition $M_{\mu\nu} = -M_{\nu\mu}$ is imposed.

Given the choice $p'_\mu = (1,0,0,1)$, the stabilizer $E(2)$ is generated by the elements

$$T_1 = B_1 - J_2 = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}; \qquad T_2 = B_2 + J_1 = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \end{pmatrix}, \tag{8.158}$$

and J_3, each of which duly annihilates $(1,0,0,1)$. The commutation relations are

$$[T_1, T_2] = 0, \qquad [J_3, T_1] = T_2, \qquad [J_3, T_2] = -T_1, \tag{8.159}$$

as appropriate for $E(2)$. Here (T_1, T_2) generate $\mathbb{R}^2$ and J_3 generates $SO(2)$ in $SO(2) \ltimes \mathbb{R}^2$.

As we have seen, the true degrees of freedom at $\mathbf{p}' = (0,0,1)$, i.e. $p'_\mu = (1,0,0,1)$ are the transverse components $\hat{A}_1$ and $\hat{A}_2$. The action of $E(2) \subset L$ on $\hat{A}_\mu(\mathbf{p}')$ given (in infinitesimal form) by the above matrices then descends to an action on $\mathbb{C}^2$ as realized by $(\hat{A}_1, \hat{A}_2)$, seen as the quotient of $\mathbb{C}^4$, consisting of the 4-vectors $(\hat{A}_0, \hat{A}_1, \hat{A}_2, \hat{A}_3)$ with $\hat{A}_0 = \hat{A}_3$, by the $\mathbb{C}$-action

$$(\hat{A}_3, \hat{A}_1, \hat{A}_2, \hat{A}_3) \mapsto (\hat{A}_3 + i\lambda, \hat{A}_1, \hat{A}_2, \hat{A}_3 + i\lambda). \tag{8.160}$$

From the above matrices, this gives

$$T_1 \begin{pmatrix} \hat{A}_1 \\ \hat{A}_2 \end{pmatrix} = 0; \qquad T_2 \begin{pmatrix} \hat{A}_1 \\ \hat{A}_2 \end{pmatrix} = 0; \qquad J_3 \begin{pmatrix} \hat{A}_1 \\ \hat{A}_2 \end{pmatrix} = \begin{pmatrix} \hat{A}_2 \\ -\hat{A}_1 \end{pmatrix}. \tag{8.161}$$

This means that the $\mathbb{R}^2$ in $E(2)$ acts trivially whilst the $SO(2)$ acts in its defining representation on $\mathbb{R}^2$, albeit complexified to $\mathbb{C}^2$. The hermitian matrix iJ_3 is diagonal in the basis

$$\mathbf{u}_\pm = (\mathbf{e}_1 \pm i\mathbf{e}_2)/\sqrt{2}, \tag{8.162}$$

with eigenvalues ± 1. Thus the $E(2)$-action defined in (8.149) is a direct sum of the characters u_λ with $\lambda = \pm 1$. In sum, we have proved the following result (where the ± 1 are these helicities):

Proposition 8.7 *The space obtained by solving the Maxwell equation* (8.142) *modulo the gauge transformations* (8.143) *is isomorphic, as a (Hilbert) space on which the Poincaré group P acts via* (8.141)*, to the direct sum of the unitary irreducible representations* $U_{(0,+,1)}$ *and* $U_{(0,+,-1)}$.

We now repeat this analysis for linearized GR. Instead of the vector A_μ, we have the symmetric tensor $h_{\mu\nu}$, see §8.4, which under the Poincaré group transforms as

$$(\Lambda, a) \cdot h_{\mu\nu}(x) = \Lambda_\mu{}^\alpha \Lambda_\beta{}^\alpha h_{\alpha\beta}(\Lambda^{-1}(x-a)). \tag{8.163}$$

Instead of the (free) Maxwell equation (8.142) we have

$$-\tfrac{1}{2}\Box \bar{h}_{\mu\nu} - \eta_{\mu\nu}\partial_\alpha\partial_\beta \bar{h}^{\alpha\beta} + \partial_\alpha\partial_\mu \bar{h}^\alpha_\nu + \partial_\alpha\partial_\nu \bar{h}^\alpha_\mu = 0, \tag{8.164}$$

i.e. the (free) linearized Einstein equations, see (8.116) - (8.117). Instead of the gauge transformation (8.143) we have the (linearized = infinitesimal) coordinate transformations

$$h_{\mu\nu} \mapsto h_{\mu\nu} + \partial_\mu \xi_\nu + \partial_\nu \xi_\mu. \tag{8.165}$$

Finally, instead of the Lorenz gauge condition $\partial^\mu A_\mu = 0$ we have the analogous equation

$$\partial^\mu \overline{h}_{\mu\nu} = 0, \tag{8.166}$$

see (8.115), which is obtained by linearizing the wave gauge (7.108). This reduces (8.164) to

$$\Box \overline{h}_{\mu\nu} = 0. \tag{8.167}$$

Given (8.166) - (8.167), one may carry out residual gauge transformations (8.165), provided

$$\Box \xi_\mu = 0. \tag{8.168}$$

Quite analogously to the electromagnetic case, we now show that for helicity ± 2 we have:

Proposition 8.8 *The space obtained by solving the linearized Einstein equations* (8.164) *modulo infinitesimal coordinate transformations* (8.165)*, or, equivalently, the space of solutions of the pair of equations* (8.166) - (8.167) *modulo the residual transformations* (8.165) *where* ξ_μ *satisfies* (8.168)*, is isomorphic, as a (Hilbert) space on which the Poincaré group P acts via* (8.163)*, to the direct sum of the unitary irreducible representations* $U_{(0,+,2)}$ *and* $U_{(0,+,-2)}$*.*

Proof. Once again, we start by solving (8.167) analogously to (8.146), upon which a simple computation based on (8.166) and (8.165) shows that (8.152) - (8.154) is replaced by

$$\begin{array}{llll}
\hat{h}_{01} = \hat{h}_{13}; & \hat{h}_{02} = \hat{h}_{23}; & \hat{h}_{03} = \frac{1}{2}(\hat{h}_{00} + \hat{h}_{33}); & \hat{h}_{22} = -\hat{h}_{11}; \\
\hat{h}_{00} \mapsto \hat{h}_{00} + 2i\hat{\xi}_0; & \hat{h}_{01} \mapsto \hat{h}_{01} + i\hat{\xi}_1; & \hat{h}_{02} \mapsto \hat{h}_{02} + i\hat{\xi}_2; & \hat{h}_{03} \mapsto \hat{h}_{03} + i(\hat{\xi}_0 + \hat{\xi}_3); \\
\hat{h}_{11} \mapsto \hat{h}_{11}; & \hat{h}_{12} \mapsto \hat{h}_{12}; & \hat{h}_{22} \mapsto \hat{h}_{22}; & \\
\hat{h}_{13} \mapsto \hat{h}_{13} + i\hat{\xi}_1; & \hat{h}_{23} \mapsto \hat{h}_{23} + i\hat{\xi}_2; & \hat{h}_{33} \mapsto \hat{h}_{33} + 2i\hat{\xi}_3 &
\end{array} \tag{8.169}$$

where for simplicity we omitted the argument $\mathbf{p}'$ common to all $\hat{h}_{\mu\nu}$. We conclude that the unconstrained and ungauged degrees of freedom are $(\hat{h}_{11}, \hat{h}_{12})$. Similarly to (8.161), this gives

$$T_1 \begin{pmatrix} \hat{h}_{11} \\ \hat{h}_{12} \end{pmatrix} = 0; \qquad T_2 \begin{pmatrix} \hat{h}_{11} \\ \hat{h}_{12} \end{pmatrix} = 0; \qquad J_3 \begin{pmatrix} \hat{h}_{11} \\ \hat{h}_{12} \end{pmatrix} = 2 \begin{pmatrix} \hat{h}_{12} \\ -\hat{h}_{11} \end{pmatrix}, \tag{8.170}$$

where the factor 2 arises from the product Λ-action on the right-hand side (8.163), which in turn leads to a sum of J_3-actions. So also here, we introduce polarized states (8.162), where this time $\mathbf{e}_i$ is the unit vector for the $\hat{h}_{1i}$ component, where $i = 1, 2$. □

From the point of view of the representation theory of the Poincaré group, linearized gravity describes massless particles (gravitons) with helicity ± 2. The full Einstein equations then add self-interactions of this particle in a seemingly beautiful and consistent way. Unfortunately, no one has been able so far to construct a renormalizable quantum field theory on this basis. But this argument does show that the origin of the diffeomorphism invariance of GR, though here just represented at some infinitesimal or linearized level, may have it origins in quantum theory, in being a space-time covariant way to describe a certain massless representation of the Poincaré group, which is related to orbits in momentum space and realized on Hilbert space.[374]

[374] One may criticize this approach for being a hybrid between classical and quantum reasoning, the former on the side of the linearized Einstein field equations and the latter on the side of the unitary representation theory of the Poincaré group on Hilbert space. But in fact a similar argument may be set up in a completely classical context, where the role of unitary representations is replaced by that of coadjoint orbits. The argument even improves, since implementing the gauge condition and quotienting by the action of the gauge group is unified into the single procedure of Marsden–Weinstein reduction. See Landsman & Wiedemann (1995) and Landsman (1998).

8.6 Conformal analysis of the constraints

The initial value constraints (8.65) - (8.67) may be analyzed from a PDE point of view.[375] In the simplest case the metric is *static*, which means that (M,g) has a timelike Killing vector field u^μ *and* has a foliation $M = \sqcup_t \Sigma_t$ whose leaves Σ_t are orthogonal to u^μ (equivalently, $\omega_{\mu\nu} = 0$. See §8.4. In that case, in the "right" (i.e. adapted) coordinates the $g_{\mu\nu}$ are time-independent, as for the Minkowski metric or the Schwarzschild solution. Hence $\tilde{k} = 0$, and if we also assume vacuum for simplicity, then the only constraint on the ensuing initial data $(\Sigma, \tilde{g}_{ij})$ is

$$\tilde{R} = 0. \tag{8.171}$$

This is a vastly underdetermined system, since the six independent components of the metric $\tilde{g}_{ij}$ are subject to just one equation. But this doesn't mean that the solution is trivial, and in particular one should understand the degrees of freedom. This is a problem in pure Riemannian geometry, whose solution as sketched below has a long and interesting history, which is worth recalling.

This history goes back to the ***uniformization theorem*** for Riemann surfaces:[376]

Theorem 8.9 *A simply connected Riemann surface is biholomorphically equivalent to one of:*

- *The Riemann sphere* $\mathbb{S}$*;*
- *The complex plane* $\mathbb{C}$*;*
- *The upper half plane* $\mathbb{H}$.

Consequently, any compact Riemann surface Σ *is (biholomorphically) isomorphic to* $\mathbb{U}/\Gamma$*, where* $\mathbb{U}$ *is* $\mathbb{S}$*,* $\mathbb{C}$*, or* $\mathbb{H}$*, and* Γ *is a discrete subgroup of the group of biholomorphic bijections of* $\mathbb{U}$ *acting freely and discontinuously on* $\mathbb{U}$ *(i.e., no* Γ*-orbit has an accumulation point).*[377]

This is equivalent to the following statement purely in the language of Riemannian geometry:

Theorem 8.10 *A complete Riemannian metric on a simply connected* 2*d manifold (and hence on a compact* 2*d Riemannian manifold) is conformally equivalent to a metric with constant curvature, cf. Theorem 4.9 (from which compact spaces may be constructed as in Theorem 8.9).*

Inspired by Theorem 8.10, the ***Yamabe problem*** asks if in arbitrary dimension any complete Riemannian metric on some closed manifold is conformally equivalent to a metric with constant Ricci scalar.[378] This problem has been solved in the positive for *compact* manifolds (which are automatically complete), using the following strategy.[379] In $d = 3$, rescale the metric by

$$\tilde{g} = \Omega^4 \gamma, \tag{8.172}$$

[375] The approach in this section goes back to Racine (1934) and Lichnerowicz (1944, 1957). For further developments see Choquet-Bruhat & York (1980), Bartnik & Isenberg (2004), Chruściel (2010), Chruściel, Galloway, & Pollack (2010), Corvino & Pollack (2011), Isenberg (2014), Galloway, Miao, & Schoen (2015), and Rácz (2015).

[376] For a historical survey of the uniformization theorem see de Saint-Gervais (2010). Jones & Singerman (1987) is an accessible low-key treatment. A Riemann *surface* is defined through its complex structure, whereas a Riemannian *manifold* is defined by its metric. In dimension 2, complex structures up to biholomorphic equivalence bijectively correspond to Riemannian metrics up to the equivalence relation defined by isometry and conformal equivalence. See also footnote 486. By (our) convention, a simply connected space is also connected.

[377] Equivalently, each $x \in \mathbb{U}$ has a nbhd U such that $U \cap \gamma \cdot U = \emptyset$ for all $\gamma \neq e$.

[378] This is the only choice among the many equivalent notions of curvature (which all coincide in $d = 2$) for which there is any hope for the problem to have a solution. See §4.5.

[379] The solution is due to Schoen (1984). See Lee & Parker (1987) and Bär (2007/08) for complete treatments.

where the conformal factor $\Omega \in C^\infty(\Sigma)$ is strictly positive (so that $\tilde{g}$ is a Riemannian metric on Σ), such that the Ricci scalar $\tilde{R} = \tilde{R}_{\tilde{g}}$ of $\tilde{g}$ is constant.[380] Straightforward computations give

$$\tilde{R} = -8\Omega^{-5} L_\gamma \Omega, \tag{8.173}$$

where the linear differential operator L_γ, called the ***conformal Laplacian***,[381] is given by

$$L_\gamma := \Delta_\gamma - \tfrac{1}{8} R_\gamma, \tag{8.174}$$

in which $\Delta_\gamma := \gamma^{ij}\nabla_i\nabla_j$ is the Laplacian on Σ defined by γ, and R_γ is the Ricci scalar defined by γ (we omit tildes on geometric quantities defined by γ; those with a tilde are defined by $\tilde{g}$). Given γ, the constraint (8.171) then becomes an equation for the scalar Ω, namely

$$L_\gamma \Omega = 0. \tag{8.175}$$

This is a linear elliptic PDE, which can indeed be solved if Σ is compact. In GR, this argument applies more generally (e.g. assuming $\Omega \to 0$ at infinity in the non-compact case).

Ellipticity is here to stay, but linearity is typical of the assumption $\tilde{k} = 0$, and in general will be replaced by gruesome nonlinearities. Indeed, already the next case, where

$$\tilde{k}_{ij} \neq 0; \qquad \mathrm{Tr}\,(\tilde{k}) := \tilde{g}^{ij}\tilde{k}_{ij} = 0, \tag{8.176}$$

is highly nonlinear.[382] The constraints (8.65) - (8.67), again in the vacuum case, simplify to

$$\tilde{R} - \mathrm{Tr}\,(\tilde{k}^2) = 0; \tag{8.177}$$

$$\tilde{g}^{jl}\tilde{\nabla}_l \tilde{k}_{ij} = 0. \tag{8.178}$$

We now also choose some symmetric tensor k_{ij} on Σ, such that

$$\gamma^{jl}\nabla_l k_{ij} = 0, \tag{8.179}$$

but freely otherwise. It is easy to show that if we relate $\tilde{k}$ to k via

$$\tilde{k}_{ij} = \Omega^{-2} k_{ij}, \tag{8.180}$$

then (8.179) implies (8.178) and hence only (8.177) remains, which is equivalent to

$$L_\gamma \Omega + \tfrac{1}{8}\mathrm{Tr}\,(k^2)\Omega^{-7} = 0. \tag{8.181}$$

This equation can be analyzed by traditional methods from nonlinear elliptic PDEs (notably by constructing both sub- and super-solutions, i.e. replacing "$= 0$" by "≤ 0" and "≥ 0").

[380] In the context of GR, adding a cosmological constant λ modifies (8.171) to $\tilde{R} = 2\lambda$. The possible signs of $\tilde{R}$, i.e. $\tilde{R} = 0, \pm 1$ up to rescaling, are restricted by the topology of Σ and define the so-called ***Yamabe class*** of Σ.

[381] Our formulae are for $d = 3$. In general dimension d, eq. (8.172) should be $\tilde{g} = \Omega^{4/(d-2)}\gamma$, whilst the conformal Laplacian is $L_\gamma = \Delta_\gamma - c_d R_\gamma$, with $c_d := \frac{1}{4}(d-2)/(d-1)$. Then (8.173) reads $\tilde{R} = -(c_d \Omega^{(d+2)/(d-2)})^{-1} L_\gamma \Omega$.

[382] Foliations with $\mathrm{Tr}\,(\tilde{k}) = 0$ are called ***maximal slicings***. This is related to the ***Plateau Problem***: if $\Sigma \subset M$ has $\mathrm{Tr}\,(\tilde{k}) = 0$, and $\mathscr{S} \subset \Sigma$ is a two-dimensional surface, then the volume of any three-dimensional $S \subset \Sigma$ with $\partial S = \mathscr{S}$ is maximal compared to the volume of competing $S \subset M$ subject to $\partial S = \mathscr{S} \subset \Sigma$. In the purely Riemannian Plateau Problem the volume (or, as in the original problem in one dimension lower, the surface area of the enclosed region) would be minimal, but in the Lorentzian case it is maximal, for similar reasons why the length of timelike geodesics is maximal rather than minimal (see §5.6): excursions of S outside Σ are in the timelike direction and hence, through lightlike approximations, *reduce* the volume (rather than increase it as in the Plateau Problem). See also §10.11.

We move to the general case. Here is it customary and physically relevant to move to a ***transverse traceless*** version of k and $\tilde{k}$, where the traceless part is easy to define, namely

$$\tilde{\sigma}_{ij} = \tilde{k}_{ij} - \tfrac{1}{3}\mathrm{Tr}\,(\tilde{k})\tilde{g}_{ij}; \qquad \sigma_{ij} = k_{ij} - \tfrac{1}{3}\mathrm{Tr}\,(k)\gamma_{ij}. \tag{8.182}$$

Adding energy-momentum and using the scaling (8.180), this reformulates the constraints as

$$L_\gamma\Omega + \tfrac{1}{8}\mathrm{Tr}\,(\sigma^2)\Omega^{-7} - \tfrac{1}{12}\mathrm{Tr}\,(k)^2\Omega^5 = -2\pi E\Omega^5; \tag{8.183}$$

$$\nabla_j\sigma_{ij} - \tfrac{2}{3}(\nabla_i\mathrm{Tr}\,(k))\Omega^6 = 8\pi P_i\Omega^{10}. \tag{8.184}$$

The first of these (i.e. the Hamiltonian constraint) is called the ***Lichnerowicz equation***. Defining the *transverse* part of σ and $\tilde{\sigma}$ is less straightforward: there exists a decomposition

$$\sigma_{ij} = \sigma^{\mathrm{TT}}_{ij} + (\hat{K}_\gamma X)_{ij}, \tag{8.185}$$

where $\sigma^{\mathrm{TT}}_{ij}$ is traceless and transverse in the sense that

$$\mathrm{Tr}\,(\sigma) \equiv \gamma^{ij}\sigma_{ij} = 0; \tag{8.186}$$

$$\nabla^i\sigma^{\mathrm{TT}}_{ij} = 0, \tag{8.187}$$

and X is some vector field, on which the ***conformal Killing operator*** $\hat{K}_\gamma$ acts by

$$(\hat{K}_\gamma X)_{ij} = \nabla_i X_j + \nabla_j X_i - \tfrac{2}{3}\gamma_{ij}\nabla_k X^k. \tag{8.188}$$

This generalizes the usual Killing operator

$$K_\gamma X = \nabla_i X_j + \nabla_j X_i, \tag{8.189}$$

whose solutions $K_\gamma X = 0$ are vector fields whose flow φ_t consists of isometries, i.e., $\varphi_t^*\gamma = \gamma$; vector fields solving $\hat{K}_\gamma X = 0$ are vector fields whose flow φ_t consists of *conformal isometries*, in that $\varphi_t^*\gamma = \Omega\gamma$ for some $\Omega > 0$, as above. The difficult part is the reconstruction of σ_{ij} from its transverse traceless part $\sigma^{\mathrm{TT}}_{ij}$ and X. This may be done by solving a conformal version of the Laplace equation, viz.

$$\hat{\Delta}_\gamma X^i = \nabla_j(\hat{K}_\gamma X)^{ij} = \Delta X^i + \tfrac{1}{3}\nabla^i\nabla_j X^j + R^i_j X^j. \tag{8.190}$$

Note that the kernel of $\hat{\Delta}_\gamma$ consists of conformal Killing vectors. Likewise for $\tilde{g}$ and $\tilde{\sigma}_{ij}$. In terms of the *free data* γ_{ij}, $\sigma^{\mathrm{TT}}_{ij}$, and $\tau \equiv \mathrm{Tr}\,(k)$, the *determined data* Ω and X are found by solving the final (conformal) version of the constraints, namely

$$L_\gamma\Omega + \tfrac{1}{8}\mathrm{Tr}\,(\sigma^2_{\mathrm{TT}})\Omega^{-7} - \tfrac{1}{12}\tau^2\Omega^5 = -2\pi E\Omega^5; \tag{8.191}$$

$$\hat{\Delta}_\gamma X^i - \tfrac{2}{3}(\nabla_i\tau)\Omega^6 = 8\pi P_i\Omega^{10}. \tag{8.192}$$

Once this has been done, $\tilde{g}_{ij}$ and $\tilde{k}_{ij}$ can be (re)constructed via (8.172) and

$$\tilde{k}_{ij} = (\hat{K}_\gamma X_{ij} + \sigma^{\mathrm{TT}}_{ij})\Omega^{-10} + \tfrac{1}{3}\tau\Omega^{-4}\gamma_{ij}, \tag{8.193}$$

and these solve the original constraints (8.65) - (8.67) in terms of the above free data. Of course, the solvability of (8.191) - (8.192) is a difficult matter, which so far is only under complete control if $\mathrm{Tr}\,(\tilde{k}) = 0$. In general the cases where Σ is compact or asymptotically flat are very different, as usual in the initial-value approach to GR, and the field is still in development.[383]

[383] See e.g. Galloway, Miao, & Schoen (2015), Holst, Maxwell, & Mazzeo (2017), and Carlotto (2021).

8.7 Hamiltonian formulation of general relativity

The Einstein equations admit a (constrained) Hamiltonian formulation, which goes back to Dirac and (independently) Bergmann in the 1950s. Their work was streamlined by Arnowitt, Deser, and Misner in the early 1960s, and in the 1970s was brought into mathematically rigorous form by various teams.[384] The Hamiltonian approach does not differ dramatically from the PDE approach as presented in §7.6 and §8.3, where both the initial data $(\Sigma, \tilde{g}, \tilde{k})$ and the equations of motion (8.61) - (8.62) were already brought into an almost Hamiltonian form, except that the Hamiltonian and the Poisson brackets were missing. The original motivation for the Hamiltonian formalism, namely to provide a basis for ("canonical") quantum gravity, remains to be fulfilled,[385] but even at the classical level it is very useful–though not indispensible–for treating boundary terms, symmetries, and conserved quantities (see below in this section as well as §8.9).

As in the previous $3+1$ first-order version of the initial value problem of GR, also in the Hamiltonian formalism the role of general covariance in the original equations (or in the Einstein–Hilbert action) is replaced by the freedom of choosing a foliation of our space-time M, and once again this freedom is parametrized by the freely specifiable lapse and shift functions. Since the Hamiltonian equations turn out to be (first-order) hyperbolic and hence deterministic, all arbitrariness in the solution lies in the choice of the lapse and the shift (and hence of the foliation).

Thus we have a time function $\mathtt{t}$ and ensuing foliation (8.1), where $\Sigma_t \cong \Sigma$ for a single $3d$-manifold Σ, and each hypersurface Σ_t is assumed to be spacelike in M. The action $S(g)$, from which the Hamiltonian will be derived, is defined on $V \subset M$, for which we assume that

$$V = \sqcup_{t\in[t_i,t_f]}\Sigma'_t; \qquad\qquad \Sigma'_t = \Sigma_t \cap V, \tag{8.194}$$

so that, as a shadow of the factorization $M \cong \Sigma \times \mathbb{R}$, we have

$$V \cong \Sigma' \times [t_i, t_f], \tag{8.195}$$

where $\Sigma' \cong \Sigma'_t$. If Σ and hence each Σ_t is closed (= compact without boundary) we assume that $\Sigma'_t = \Sigma_t$; otherwise (think of $\Sigma \cong \mathbb{R}^3$), $\Sigma'_t \subset \Sigma_t$ is a compact submanifold with boundary

$$S_t := \partial\Sigma'_t \tag{8.196}$$

in Σ_t (think of $\Sigma'_t \cong B^3_r$, the closed 3-ball in $\mathbb{R}^3$ with radius r, so that $S_t \cong \partial B^3_r = S^2_r$, the 2-sphere in $\mathbb{R}^3$ with radius r). This means that the boundary ∂V of V decomposes as

$$\partial V = \Sigma_{t_i} \cup \Sigma_{t_f} \cup B; \qquad\qquad B = \cup_{t\in[t_i,t_f]} S_t, \tag{8.197}$$

which is a (hyper)cylinder bounded above and below by 3-manifolds Σ_{t_f} and Σ_{t_i}, respectively, and bounded on the side by a 3-manifold B that in turn is foliated by the 2-manifolds S_t. Using

$$g = -L^2\tilde{g}; \qquad\qquad \Rightarrow \qquad\qquad \sqrt{-g} = L\sqrt{\tilde{g}}, \tag{8.198}$$

[384] Pioneering papers include Dirac (1950, 1958ab), Bergmann (1949), Bergmann & Brunings (1949), and Arnowitt, Deser, & Misner (1962), whose approach is reviewed in Misner, Thorne, & Wheeler (1973), §21.6. See Salisbury (2020) for the history of canonical GR. Of the reviews in the theoretical physics literature we mention Poisson (2004), §4.2, and Sundermeyer (2014), chapter 7 and §C.3. The mathematics was done, in different ways, by e.g. Fischer & Marsden (1979), Kijowski & Tulczyjew (1979), and Isenberg & Nester (1980). Dirac's approach, which is still used, involved a heavy "constraint algorithm", which can be avoided as long as one realizes that the ultimate justification of any Hamiltonian formalism is that it simply reproduces the Einstein equations.

[385] See DeWitt (1967), Rovelli (2004), and Thiemann (2007). As for nuclear fusion, one begins to lose patience.

where $g \equiv \det g$ and $\tilde{g} \equiv \det \tilde{g}$, which follows from (8.14), as well as (8.58), we may then rewrite the Einstein–Hilbert action (7.2) and the boundary term (7.36). This gives

$$S_G(g) = \int_{t_i}^{t_f} dt \int_{\Sigma_t} d^3y \sqrt{\tilde{g}_t(y)}\,[L(\tilde{R} + \mathrm{Tr}\,(\tilde{k})^2 + \mathrm{Tr}\,(\tilde{k}^2)) - 2(\mathscr{L}_{e_0}\mathrm{Tr}\,(\tilde{k}) + \tilde{\Delta} L)]; \tag{8.199}$$

$$S_B(g) = 2\int_{\Sigma_{t_f}} d^3y \sqrt{|\det(\tilde{g})|}\,\mathrm{Tr}\,(\tilde{k}_t) - 2\int_{\Sigma_{t_i}} d^3y \sqrt{|\det(\tilde{g})|}\,\mathrm{Tr}\,(\tilde{k}_t)$$
$$- 2\int_{t_i}^{t_f} dt \int_{S_t} d^2z \sqrt{|\det(\hat{g})|}\,\mathrm{Tr}\,(\hat{k}_t)) - \cdots, \tag{8.200}$$

where $\tilde{g}_t$ and $\tilde{k}_t$ are the induced 3-metric and the extrinsic curvature on Σ_t with regard to its embedding $\Sigma_t \hookrightarrow V \subset M$, respectively, and likewise $\hat{k}_t$ is the extrinsic curvature on $B \hookrightarrow V$. The dots in (8.200) mean that for the moment we omit the $\tilde{k}^0$ terms in (7.36), which will be reinstalled at the end of the calculation. The last term in (8.199) is also a boundary term, since

$$\int_{\Sigma_t} d^3y \sqrt{\tilde{g}_t(y)}\tilde{\Delta} L = \int_{\Sigma} d^y x\, \partial_i(\sqrt{\tilde{g}_t(y)}\tilde{\nabla}^i L) = \int_{S_t} d^2\vec{\sigma}^i \cdot \tilde{\nabla}_i L, \tag{8.201}$$

similarly to (7.17) - (7.18), but now in $3d$. Using (8.10), the penultimate term in (8.199) equals

$$\mathscr{L}_{e_0}\mathrm{Tr}\,(\tilde{k}) = e_0\mathrm{Tr}\,(\tilde{k}) = LN^\mu \partial_\mu \mathrm{Tr}\,(\tilde{k}) = L(\nabla_\mu(\mathrm{Tr}\,(\tilde{k})N^\mu) - \mathrm{Tr}\,(\tilde{k})\nabla_\mu N^\mu). \tag{8.202}$$

The first term gives a boundary term that cancels the first two terms in (8.200). The second is

$$\nabla_\mu N^\mu = -\mathrm{Tr}\,(\tilde{k}), \tag{8.203}$$

as follows from (8.40), in which $\tilde{\partial}_\nu$ is spatial, so that $N^\mu \tilde{\partial}_\nu = 0$. Finally, (8.201), which came from the bulk action (8.199), and the last term in the boundary action (8.200) neatly combine to

$$\int_{S_t} d^2z \sqrt{|\det(\hat{g})|}\,(\vec{n}^i\tilde{\nabla}_i L + \mathrm{Tr}\,(\hat{k}_t)) = \int_{S_t} d^2z \sqrt{|\det(\breve{g}_t)}\,L\,\mathrm{Tr}\,(\breve{k}_t), \tag{8.204}$$

where $\vec{n}$ is the outward normal vector into Σ_t of the embedding $S_t \hookrightarrow \Sigma_t$, $\breve{g}_t$ is the induced metric on the 2-manifold S_t, and $\breve{k}_t$ is the extrinsic curvature on Σ_t, with respect to the embedding $S_t \hookrightarrow \Sigma_t$.[386] Combining all, and reinserting the constant $\tilde{k}^0$ terms in (7.36), gives the total action

$$S(g) = S_G(g) + S_B(g) = \int_{t_i}^{t_f} dt \left(\int_{\Sigma_t} d^3y \sqrt{\det(\tilde{g}_t(y))}\,L \cdot (\tilde{R} + \mathrm{Tr}\,(\tilde{k}_t^2) - \mathrm{Tr}\,(\tilde{k}_t)^2) \right.$$
$$\left. - 2\int_{S_t} d^2z \sqrt{\det(\breve{g}_t(z))}\,L\,(\mathrm{Tr}\,(\breve{k}_t) - \mathrm{Tr}\,(\breve{k}_t^0)) \right), \tag{8.205}$$

where $\breve{k}_t^0$ is the extrinsic curvature of the embedding of the surface S_t in $\mathbb{R}^3$, seen as the $x^0 = t$ slice of Minkowski space-time. We repeat that this term is necessary for convergence if Σ_t approaches Σ, in case that Σ is not compact (if Σ is closed all boundary terms may be ignored).

In order to pass from a Lagrangian to a Hamiltonian description, we write, with $x = (y,t)$,

$$S_G(g) = \int_{t_i}^{t_f} dt\, \mathscr{L}_G(x); \tag{8.206}$$

$$\mathscr{L}_G(x) = \sqrt{\tilde{g}_t(y)}\,L(x)\,(\tilde{R}(x) + \mathrm{Tr}\,(\tilde{k}_t(y)^2) - \mathrm{Tr}\,(\tilde{k}_t(y))^2), \tag{8.207}$$

[386] See Poisson (2004), §4.2.5, for a calculation yielding (8.204); his book contains many computations omitted here. This may also be inferred by choosing $L = 1$ for the moment and noting that (8.204) is a geometric expression. Note that Poisson (and many physicists) defines the extrinsic curvature as minus ours (the math convention).

in which $\tilde{k}$ (and hence the quantities $\mathrm{Tr}\,(\tilde{k})$ and $\mathrm{Tr}\,(\tilde{k}^2)$ derived from it) is determined by (8.22). If we use (8.64) in the latter expression, in terms of spatial indices, we obtain

$$\tilde{k}_{ij} = \tfrac{1}{2}L^{-1}(\tilde{g}_{ik}\tilde{\nabla}_j S^k + \tilde{g}_{jk}\tilde{\nabla}_i S^k - \partial_t \tilde{g}_{ij}), \tag{8.208}$$

which allows us to compute the canonical momenta for the 3-metric $\tilde{g}_{ij}$, as in $p_i = \partial L/\partial \dot{q}^i$, viz.

$$\tilde{\pi}^{ij} := \frac{\partial \mathscr{L}}{\partial \dot{\tilde{g}}_{ij}} = \sqrt{\tilde{g}}(\mathrm{Tr}\,(\tilde{k})\tilde{g}^{ij} - \tilde{k}^{ij}). \tag{8.209}$$

Thus the $\tilde{k}_{ij}$ will be seen as functions of $\tilde{g}_{ij}$ and $\tilde{\pi}^{ij}$ by inverting (8.209), which yields

$$\sqrt{\tilde{g}}\tilde{k}_{ij} = \tfrac{1}{2}\mathrm{Tr}\,(\tilde{\pi})\tilde{g}_{ij} - \tilde{\pi}_{ij}, \tag{8.210}$$

where $\mathrm{Tr}\,(\tilde{\pi}) = \tilde{g}_{kl}\tilde{\pi}^{kl}$ and $\tilde{\pi}_{ij} = \tilde{g}_{ik}\tilde{g}_{jl}\tilde{\pi}^{kl}$. One also uses the 'de-densitized' momentum[387]

$$\check{\pi}_{ij} = \tilde{\pi}_{ij}/\sqrt{\tilde{g}} = \mathrm{Tr}\,(\tilde{k})\tilde{g}_{ij} - \tilde{k}_{ij}, \tag{8.211}$$

so that

$$\tilde{k}_{ij} = \tfrac{1}{2}\mathrm{Tr}\,(\check{\pi})\tilde{g}_{ij} - \check{\pi}_{ij}. \tag{8.212}$$

Note that the boundary action $S_B(g)$ does not contain $\dot{\tilde{g}}_{ij}$ and hence makes no contribution to $\tilde{\pi}^{ij}$. Furthermore, since neither S_G nor S_B contains the time derivatives $\dot{L}$ and $\dot{S}^i$ of the lapse and the shift, the corresponding momenta vanish and may be ignored. The canonical Hamiltonian

$$H(p_i, q^i) = \sum_i p_i \dot{q}^i - L(q^i, \dot{q}^i), \tag{8.213}$$

where the $\dot{q}^i$ are to be eliminated in favour of their conjugate momenta, may then be computed as usual. For GR this gives two terms, coming from the bulk and the boundary Lagrangians. First,

$$H'_G = \int_\Sigma d^3y\, H'_G(y) = \lim_{\Sigma' \nearrow \Sigma} H'_G(\Sigma') = \lim_{\Sigma' \nearrow \Sigma} \int_{\Sigma'} d^3y\, H'_G(y); \tag{8.214}$$

$$H'_G(\tilde{\pi}^{ij}, \tilde{g}_{ij}) = \tilde{\pi}^{ij}\partial_t \tilde{g}_{ij} - \mathscr{L}_G(\tilde{g}_{ij}, \partial_t \tilde{g}_{ij}), \tag{8.215}$$

where in the spirit of the Hamiltonian formalism (called "geometrodynamics" in this regard), we have replaced Σ_t by Σ and hence regard $\tilde{g}_{ij}$ and $\check{\pi}_{ij}$ as (geometric) quantities defined on Σ. However, although S_B does not contribute to the definition of the momenta, it plays the role of a potential energy, and as such should be (negatively) added to the total Hamiltonian, in that

$$H(\Sigma') = H'_G(\Sigma') + H'_B(\Sigma'); \tag{8.216}$$

$$H'_B(\Sigma') = 2\int_{\partial\Sigma'} d^2z \sqrt{\det(\check{g}_t(z))}\, L\,(\mathrm{Tr}\,(\check{k}) - \mathrm{Tr}\,(\check{k}^0)), \tag{8.217}$$

cf. (8.205). We wrote primes here, because, like the original bulk action S_G, the bulk Hamiltonian H'_G in fact contains divergences leading to boundary terms. Indeed, if we solve (8.208) for $\partial_t \tilde{g}_{ij}$,

[387] As before, define a volume $\omega \in \Omega^3(\Sigma)$ by $\omega_x = \sqrt{\tilde{g}(x)}dx^1 \wedge dx^2 \wedge dx^3$. Geometrically, the canonical momentum $\tilde{\pi}$ conjugate to the spatial metric $\tilde{g}$ should be regarded as an element of $\mathfrak{X}^{(0,2)}(\Sigma) \otimes \Omega^3(\Sigma)$, on which interpretation (8.209) should be written as $\tilde{\pi} = (\mathrm{Tr}\,(\tilde{k})\tilde{g} - \tilde{k})^\sharp \otimes \omega$, or $\check{\pi}^{ij} = \sqrt{\tilde{g}}(\mathrm{Tr}\,(\tilde{k})\tilde{g}^{ij} - \tilde{k}^{ij})dx^1 \wedge dx^2 \wedge dx^3$.

and substitute this in (8.215), through a partial integration and Stokes's theorem (on Σ, i.e. in $3d$) one may replace the terms involving $\tilde{\nabla}_j S^k$ and $\tilde{\nabla}_i S^k$ by terms linear in S^k. This yields

$$H(\Sigma) = H_G(\Sigma) + H_B(\Sigma) = \lim_{\Sigma' \nearrow \Sigma} H_G(\Sigma') + H_B(\Sigma'); \tag{8.218}$$

$$H_G(\Sigma') = \int_{\Sigma'} d^3y \sqrt{\det(\tilde{g}(y))} H_G(y); \tag{8.219}$$

$$H_B(\Sigma') = \int_{\partial\Sigma'} d^2z \sqrt{\det(\check{g}(z))} H_B(z); \tag{8.220}$$

where the true bulk and boundary Hamiltonian densities are given by

$$H_G = LC_0 + S^i C_i; \tag{8.221}$$

$$H_B = 2(L(\mathrm{Tr}(\check{k}) - \mathrm{Tr}(\check{k}^0)) + S^i \vec{n}^j \check{\pi}_{ij}). \tag{8.222}$$

These, in turn, are defined in terms of the familiar expressions, cf. (7.148) - (7.149),

$$C_0 = -\tilde{R} + \mathrm{Tr}(\tilde{k}^2) - \mathrm{Tr}(\tilde{k})^2 = -\tilde{R} + \mathrm{Tr}(\check{\pi}^2) - \tfrac{1}{2}\mathrm{Tr}(\check{\pi})^2; \tag{8.223}$$

$$C_i = -2(\tilde{\nabla}_j \tilde{k}^j_i - \tilde{\nabla}_i \mathrm{Tr}(\tilde{k})) = -2\tilde{\nabla}_j \tilde{\pi}^j_i. \tag{8.224}$$

Here the lapse and the shift are (freely) given functions of space and time, whereas the canonical quantities $(\tilde{g}_{ij}, \tilde{\pi}^{ij})$ or, equivalently, $(\tilde{g}_{ij}, \check{\pi}^{ij})$, which are initially defined on Σ, evolve according to the Hamiltonian equations of motion, and as such become time-dependent (what this "time" means becomes clear only when the total globally hyperbolic space-time (M, g) plus its foliation dictated by the solution is reconstructed). If Σ is compact one may put $\Sigma' = \Sigma$ and forget about the boundary terms (i.e. $\partial\Sigma' = \partial\Sigma = \emptyset$). If Σ is non-compact and each approximant $\Sigma' \subset \Sigma$ is compact, the metric and extrinsic curvature $(\check{g}, \check{k})$ of $\partial\Sigma'$ seen as embedded in Σ' are determined by $(\tilde{g}, \check{\pi})$ are hence are dependent variables. Moreover, since C_0 and C_i are the Hamiltonian and momentum constraints (7.148) - (7.149), respectively, we need to put

$$C_0 = 0; \qquad\qquad C_i = 0. \tag{8.225}$$

We may simply do this "by hand", since as mentioned before, the ultimate justification of any Hamiltonian formulation should lie in its equivalence with the original Lagrangian formulation of the problem.[388] One could also treat (L, S^i) as canonical variables, and notice that, because the action (8.199) - (8.200) does not contain their time derivatives, the associated canonical momenta vanish. Using the Hamiltonian (8.218), from the Hamiltonian equations of motion

$$\dot{q}^i = \partial H / \partial p_i; \qquad\qquad \dot{p}_i = -\partial H / \partial q^i \tag{8.226}$$

we therefore obtain something like $\partial H / \partial L = -\dot{\tilde{\pi}}_L = 0$. This gives $C_0 = 0$, noting that the variation of L (like that of $\tilde{g}_{ij}$, but not that of $\tilde{\pi}_{ij}$) is supposed to vanish on the boundary, so that $H_B(\Sigma')$ makes no contribution to the equations of motion (although it is a crucial part of the Hamiltonian itself, as we shall see). Similarly, $\partial H / \partial S^i = 0$ gives the spatial constraint $C_i = 0$.

The real equations of motion come from (8.226) applied to $\tilde{g}_{ij}$ and $\tilde{\pi}^{ij}$, as follows.[389]

[388] If $(\Sigma, \tilde{g})$ is asymptotically flat, flor $L = 1$ and $S^i = 0$ the boundary Hamiltonian $H_B(\Sigma)$, which on the solution to the constraints (8.225) *is* the Hamiltonian, equals the total mass (8.126). See Poisson (2004), Problem 4.6.7

[389] All boundary terms cancel, as in the Lagrangian approach, so one may as well ignore them here.

- Using (8.212), the equation $\partial_t \tilde{g}_{ij} = \partial H / \partial \tilde{\pi}^{ij}$ may be shown to coincide with (8.22), i.e.,

$$\frac{\partial \tilde{g}_{ij}}{\partial t} = 2L(\check{\pi}_{ij} - \tfrac{1}{2}\mathrm{Tr}\,(\check{\pi})\tilde{g}_{ij}) + \mathscr{L}_S \tilde{g}_{ij}. \tag{8.227}$$

- The equation $\partial_t \tilde{\pi}^{ij} = -\partial H / \partial \tilde{g}_{ij}$ takes slightly more effort to make explicit; the result is

$$\begin{aligned}\frac{\partial \tilde{\pi}^{ij}}{\partial t} &= L(\mathrm{Tr}\,(\check{\pi})\check{\pi}^{ij} - 2\check{\pi}^{ik}\check{\pi}_k^j + \tfrac{1}{2}(\mathrm{Tr}\,(\check{\pi}^2) - \tfrac{1}{2}\mathrm{Tr}\,(\check{\pi})^2)\tilde{g}^{ij} - \tilde{G}^{ij})\sqrt{\tilde{g}} \\ &\quad + (\tilde{\nabla}^i\tilde{\nabla}^j L - \tilde{g}^{ij}\tilde{g}^{kl}\tilde{\nabla}_k\tilde{\nabla}_l L)\sqrt{\tilde{g}} + \mathscr{L}_S\tilde{\pi}^{ij},\end{aligned} \tag{8.228}$$

  where $\tilde{G}_{ij} = \tilde{R}_{ij} - \frac{1}{2}\tilde{g}_{ij}\tilde{R}$ is the $3d$ Einstein tensor defined by $\tilde{g}$. Eq. (8.228) is equivalent to (8.59), and so the pair (8.227) - (8.228) is equivalent to the pair (8.61) - (8.62).

- In particular, for lapse $L = 1$ and shift $S^i = 0$ one obtains the ***Einstein flow*** equations

$$\frac{\partial \tilde{g}_{ij}}{\partial t} = 2\check{\pi}_{ij} - \mathrm{Tr}\,(\check{\pi})\tilde{g}_{ij}; \tag{8.229}$$

$$\frac{1}{\sqrt{\tilde{g}}}\frac{\partial \tilde{\pi}^{ij}}{\partial t} = \mathrm{Tr}\,(\check{\pi})\check{\pi}^{ij} - 2\check{\pi}^{ik}\check{\pi}_k^j + \tfrac{1}{2}(\mathrm{Tr}\,(\check{\pi}^2) - \tfrac{1}{2}\mathrm{Tr}\,(\check{\pi})^2)\tilde{g}^{ij} - \tilde{G}^{ij}. \tag{8.230}$$

The structure of (8.221) may be further clarified by comparison with electromagnetism (cf. §7.4). In electrodynamics (for simplicity in Minkowski space-time), the Lagrangian density is

$$\mathscr{L} = -\tfrac{1}{4}F_{\mu\nu}F^{\mu\nu}, \tag{8.231}$$

cf. (7.83), seen as a functional of A_μ. The canonical momentum

$$\pi^0 = \partial\mathscr{L}/\partial\dot{A}_0 \tag{8.232}$$

conjugate to A_0 vanishes, since $\mathscr{L}$ does not contain $\dot{A}_0$. The one conjugate to A_i equals

$$\pi^i = \partial\mathscr{L}/\partial\dot{A}_i = F^{0i} = -E^i, \tag{8.233}$$

in terms of which the Hamiltonian is

$$H = \int d^3x\,(-\vec{E}\cdot\vec{A} - \mathscr{L}(A_0, \vec{A}, \partial_t\vec{A})) = \int d^3x\,(\tfrac{1}{2}\vec{E}\cdot\vec{E} + \tfrac{1}{2}\vec{B}\cdot\vec{B} + A_0\nabla\cdot\vec{E}), \tag{8.234}$$

where $\vec{B} = \nabla\times\vec{A}$, and the term relevant to us, $A_0\nabla\cdot\vec{E}$, comes from partially integrating $-\vec{E}\cdot\nabla A_0$. Thus A_0 is like the lapse and its equation of motion gives the (Gauss law) constraint

$$\nabla\cdot\vec{E} = 0. \tag{8.235}$$

The equation of motion for A_i, i.e. $\dot{A}_i = \partial H/\partial\pi^i = -\partial H/\partial E^i = -E_i + \nabla A_0$ gives

$$\partial_t\vec{B} = -\nabla\times\vec{E}, \tag{8.236}$$

whereas the one for π^i, i.e., $\dot{E}^i = -\dot{\pi}^i = \partial H/\partial A_i$, yields

$$\partial_t\vec{E} = \nabla\times\vec{B}. \tag{8.237}$$

Maxwell's equations (*in vacuo*) are completed by the constraint (8.235), and the identity

$$\nabla\cdot\vec{B} = 0, \tag{8.238}$$

which follows from the definition $\vec{B} = \nabla\times\vec{A}$.

8.8 Constraints and deformation algebra

In the 1970s an interesting perspective on the constraints (8.65) - (8.67) arose,[390] which was believed to be relevant for (canonical) quantum gravity (a hope unfulfilled so far), but which is also crying out for further mathematical and other conceptual understanding in classical GR.[391]

We return to the setting of §8.1. Let $\mathrm{Fol}(\Sigma,M,g)$ be the "space" of all spacelike foliations F of some globally hyperbolic space-time M, as defined in §8.1, and let $\mathrm{Emb}(\Sigma,M,g)$ be the "space" of spacelike embeddings $\imath:\Sigma\hookrightarrow M$. Then each foliation F defines a curve

$$c:\mathbb{R}\to\mathrm{Emb}(\Sigma,M,g);\qquad\qquad t\mapsto c_t \tag{8.239}$$

in $\mathrm{Emb}(\Sigma,M,g)$ via $c_t(x)=F(t,x)$. Although curves in $\mathrm{Emb}(\Sigma,M,g)$ do not necessarily describe foliations of M (just think of the constant curve), it is interesting to study tangent vectors to curves that do, and regard these vectors as "infinitesimal" foliations. Take a tangent vector

$$X'\in T_\imath\mathrm{Emb}(\Sigma,M,g)\qquad\qquad X'=\frac{dc_t}{dt}(t=0), \tag{8.240}$$

for some curve $t\mapsto c_t$ in $\mathrm{Emb}(\Sigma,M,g)$ with $c_0=\imath$. Given such a vector X', we define

$$X:\imath(\Sigma)\to TM;\qquad\qquad X(\imath(x)):=\frac{dc_t(x)}{dt}(t=0), \tag{8.241}$$

so that $X(\imath(x))\equiv X_{\imath(x)}\in T_{\imath(x)}M$. The bijective correspondence $X'\leftrightarrow X$ gives an isomorphism

$$T_\imath\mathrm{Emb}(\Sigma,M,g)\cong\Gamma(\imath(\Sigma),T_{\imath(\Sigma)}M) \tag{8.242}$$

of vector spaces, where $T_{\imath(\Sigma)}M$ is the restriction of TM to $\imath(\Sigma)\subset M$, seen as a vector bundle over $\imath(\Sigma)$. We may even remove the dependence on $\imath$ by further identifying

$$\Gamma(\imath(\Sigma),T_{\imath(\Sigma)}M)\cong C^\infty(\Sigma)\oplus\mathfrak{X}(\Sigma);\quad\Rightarrow\quad T_\imath(\mathrm{Emb}(\Sigma,M,g))\cong C^\infty(\Sigma)\oplus\mathfrak{X}(\Sigma), \tag{8.243}$$

where the right-hand side is seen as a vector space (on this understanding one may write $\times$ instead of $\oplus$). Namely, one has a unique future-directed normal vector field N to $\imath(\Sigma)$, normalized to

$$g(N,N)=-1, \tag{8.244}$$

as usual. Then decompose $X=\tilde{L}N+\tilde{S}$, as in (8.5), with the difference that so far the lapse $\tilde{L}$ and the shift $\tilde{S}$ are defined on $\imath(\Sigma)$ alone (see below for their extension to M).

Before proceeding, let us first sketch a simpler situation that is well understood and which one, with limited success so far, would like to mimic in GR. If we replace $\mathrm{Emb}(\Sigma,M,g)$ by the diffeomorphism group $\mathrm{Diff}(\Sigma)$ of Σ (say by replacing $M\rightsquigarrow\Sigma$), the analogue of (8.243) is

$$T_\psi\mathrm{Diff}(\Sigma)\cong\mathfrak{X}(\Sigma), \tag{8.245}$$

[390] Key papers include Teitelboim (1973), Hojman, Kuchar, & Teitelboim (1976), Kuchar (1976), and Isham & Kuchar (1985ab). See also Anderson (2007), Gomes & Shyam (2016), and Gomes & Butterfield (2020) and references therein to later literature. A mathematically rigorous version of the Poisson brackets involved in this analysis, and more generally of the entire Hamiltonian approach to GR, including the PDE side, was simultaneously and independently developed in a series of papers culminating in the review by Fischer & Marsden (1979). All of this has so far only been done for compact Cauchy surfaces Σ, so we assume this. See also Proposition 6.19.

[391] See Blohmann, Barbosa Fernandes, & Weinstein (2013), Bojowald *et al.* (2016), Blohmann & Weinstein (2018), and Głowacki (2019), all based on Lie algebroids. We refrain from specifying topologies and smoothness; the best setting seems to be *diffeology* (Iglesias-Zemmour, 2013; van der Schaaf, 2020), as in the papers just cited.

since vector fields integrate to one-parameter groups of diffeomorphisms. In particular, if we regard $\mathrm{Diff}(\Sigma)$ as an infinite-dimensional Lie group, and take $\psi = \mathrm{id}_\Sigma$ to be the identity, then $\mathfrak{X}(\Sigma)$ is the Lie algebra of $\mathrm{Diff}(\Sigma)$, whose Lie bracket, however, is *minus* the commutator.

For another way to look at this, let G be a Lie group and N some manifold. Recall that a G-action on N is a smooth map $\varphi : G \times N \to N$, written as $\varphi(\gamma, x) \equiv \varphi_\gamma(x) \equiv \gamma x$, such that $ex = x$ and $g(hx) = (gh)x$ for all $x \in M$ and $g, h \in M$. A G-action on N gives rise to a map

$$\varphi_* : \mathrm{Lie}(G) \to \mathfrak{X}(N); \qquad \varphi_*(A) f(x) = \frac{d}{dt} f(e^{-tA} x)_{|t=0}, \tag{8.246}$$

where $A \in \mathrm{Lie}(G)$, i.e. the Lie algebra of G, and $f \in C^\infty(M)$. This map can be shown to be a Lie algebra homomorphism. Taking $G = \mathrm{Diff}(\Sigma)$ and $N = \Sigma$, with the defining action, we find

$$\varphi_* : \mathfrak{X}(\Sigma) \to \mathfrak{X}(\Sigma); \qquad X \mapsto -X, \tag{8.247}$$

whose minus sign is correct: as has just been noted, the Lie bracket on $\mathfrak{X}(\Sigma) = \mathrm{Lie}(\mathrm{Diff}(\Sigma))$ is minus the commutator. Towards (8.243), another relevant construction is to take $N = V$ to be a vector space on which G acts linearly, and form the semidirect product $V \rtimes G$, with Lie bracket

$$[(v, X), (w, Y)] = (Xw - Yv, [X, Y]), \tag{8.248}$$

defined on the vector space $\mathrm{Lie}(V \rtimes G) = V \oplus \mathrm{Lie}(G)$, where Xw is the same as $\varphi_*(X)w$ as just defined, evaluated at $T_0 V \cong \mathrm{Lie}(V) \cong V$. Take $G = \mathrm{Diff}(\Sigma)$ and $N = C^\infty(\Sigma)$, where $\mathrm{Diff}(\Sigma)$ acts on $C^\infty(\Sigma)$ by pullback of its defining action on Σ. The bracket (8.248) is then defined on $C^\infty(\Sigma) \oplus \mathfrak{X}(\Sigma)$. Writing $\tilde{L} \in C^\infty(\Sigma)$ and $\tilde{S} \in \mathfrak{X}(\Sigma)$, eq. (8.248) becomes

$$[(\tilde{L}_1, \tilde{S}_1), (\tilde{L}_2, \tilde{S}_2)] = (\mathscr{L}_{\tilde{S}_1} \tilde{L}_2 - \mathscr{L}_{\tilde{S}_2} \tilde{L}_1, [\tilde{S}_1, \tilde{S}_2]), \tag{8.249}$$

where $\mathscr{L}_{\tilde{S}} \tilde{L} = \tilde{S}\tilde{L}$ is the defining action of the vector field $\tilde{S}$ on the function $\tilde{L}$, and $[\tilde{S}_1, \tilde{S}_2] = \mathscr{L}_{\tilde{S}_1} \tilde{S}_2$ is the usual commutator of vector fields, all happening on Σ. However, returning to (8.243), consider the following closely related bracket on $C^\infty(\Sigma) \oplus \mathfrak{X}(\Sigma)$, seen as $T_\imath(\mathrm{Emb}(\Sigma, M, g))$:

$$[(\tilde{L}_1, \tilde{S}_1), (\tilde{S}_1, \tilde{S}_2)]_\imath = (\mathscr{L}_{\tilde{S}_1} \tilde{L}_2 - \mathscr{L}_{\tilde{S}_2} \tilde{L}_1, [\tilde{S}_1, \tilde{S}_2] + \tilde{L}_1 \tilde{\nabla} \tilde{L}_2 - \tilde{L}_2 \tilde{\nabla} \tilde{L}_1). \tag{8.250}$$

Here Σ has been endowed with a Riemannian metric $\tilde{g} = \imath^* g$, as in the initial value formulation, and the bracket depends on this metric through its divergence operator $\tilde{\nabla} := \nabla_{\imath^* g} = \nabla_{\tilde{g}}$ (which sends functions to vector fields).[392] In physics this bracket is called the ***deformation algebra***.

As we shall see, the Poisson bracket of the constraints in GR reproduces this algebra, and hence it would be nice to understand it better, for example by seeing it as a commutator. To this end, we note that $\mathrm{Diff}(M)$ acts on the space $\mathrm{Emb}(\Sigma, M)$ of *all* embeddings $\Sigma \hookrightarrow M$ via

$$\psi(\imath) = \psi \circ \imath, \tag{8.251}$$

but this does not restrict to an action on the space $\mathrm{Emb}(\Sigma, M, g)$ of all *spacelike* embeddings. On the other hand, if $\imath(\Sigma)$ is spacelike with respect to g, then $\psi \circ \imath$ is spacelike with respect to $\psi^* g$, so that if ψ is an isometry and $\imath \in \mathrm{Emb}(\Sigma, M, g)$, then $\psi \circ \imath \in \mathrm{Emb}(\Sigma, M, g)$, and hence one does have a well-defined action of the group $\mathrm{Iso}(M, g)$ of all isometries of (M, g) on $\mathrm{Emb}(\Sigma, M, g)$.

[392] For a fixed $3d$ metric $\tilde{g}$ this is not a Lie bracket, as the Jacobi identity may fail (Blohmann *et al.*, 2013). The following construction may also be found in this paper, shadowed by an analogous discussion in coordinates in Bojowald *et al.* (2016). For these authors, this construction is just an introduction to the use of Lie algebroids.

One can do more, however, at least locally, in the sense of validity within some open nbhd U of $\imath(\Sigma)$ in M, for some fixed $\imath \in \mathrm{Emb}(\Sigma, M, g)$ at which we explore the tangent space (8.243). As in making the identification (8.243), fix some spacelike embedding $\imath : \Sigma \hookrightarrow M$ with fd normal $\tilde{N}$, so far just defined on $\imath(\Sigma)$. By (for example) the tubular neighbourhood theorem of differential geometry and the local properties of the exponential map, there exists an open nbhd

$$U \cong I \times \Sigma \tag{8.252}$$

of $\imath(\Sigma)$ in M, where $0 \in I \subset \mathbb{R}$ is an open interval, such that the timelike geodesics with tangent $\tilde{N}$ at $\imath(\Sigma)$ (and hence normal to $\imath(\Sigma)$) do not cross within U. This gives a foliation

$$U = \sqcup_{s \in I} \Sigma_s, \tag{8.253}$$

where $\Sigma_0 = \imath(\Sigma)$ and Σ_s is the set of points $\gamma^{(x)}_{\tilde{N}}(s)$, where $x \in \Sigma_0$ and $\gamma^{(x)}_{\tilde{N}}$ is the geodesic with $\gamma^{(x)}_{\tilde{N}}(0) = x$ and $\dot{\gamma}^{(x)}_{\tilde{N}}(0) = \tilde{N}_x$. We call this local foliation, which is entirely determined by $\imath$, ***canonical***.[393] It has lapse $L = 1$ and shift $S = 0$. The normal $\tilde{N}$, so far defined on Σ, extends to a vector field N on U through parallel transport along these geodesics, i.e. by solving

$$\nabla_N N = 0; \tag{8.254}$$

as we know, this preserves the normalization (8.244). The canonical foliation then arises by simply transporting Σ_0 along the flow of N.

Definition 8.11 *A vector field $X \in \mathfrak{X}(U)$ is* **Gaussian** *iff one and hence each of the following equivalent conditions is satisfied:*[394]

1. *$\mathscr{L}_X(N,Y) = 0$ for each $Y \in \mathfrak{X}(U)$, i.e. $i_N \mathscr{L}_X g = 0$. Equivalently, $N^\mu(\nabla_\mu X_\nu + \nabla_\nu X_\mu) = 0$.*
2. *The flow ψ_t of X preserves the canonical foliation near Σ just defined.*[395]

By definition, the first condition is equivalent to the following property of the flow:

$$g_{\psi_t(x)}(T_x\psi_t N_x, T_x\psi_t Y_x) = g_x(N_x, Y_x), \tag{8.255}$$

for each $x \in \Sigma$. This implies that, as announced, the flow of a Gaussian vector field maps spacelike surfaces to spacelike surfaces, at least within U, i.e. for small enough t.[396] In that sense, Gaussian vector field reside somewhere between Killing vector fields and arbitrary ones.

Proposition 8.12 *Each vector field $\tilde{X} \in \Gamma(\Sigma_0, T_{\Sigma_0}M)$, where $\Sigma_0 := \imath(\Sigma)$, has a unique Gaussian extension X to U, which, if decomposed as $X = LN + S$, has lapse L and shift S satisfying*

$$\mathscr{L}_N L = 0; \qquad\qquad \mathscr{L}_N S = \tilde{\nabla} L, \tag{8.256}$$

where $\tilde{\nabla} L = \nabla L + g(N, \nabla L)N$ is the spatial gradient of L, i.e. $\tilde{\nabla}^\mu L = (g^{\mu\nu} + N^\mu N^\nu)\partial_\nu L$. This is tangent to the leaves of the canonical foliation, so that $g(\mathscr{L}_N S, N) = g(\tilde{\nabla} L, N) = 0$.

[393] In general it cannot be extended to M since such geodesics may cross. We assume Σ is compact.

[394] The name *Gaussian* comes from the fact that the flow ψ_t also preserves the "Gaussian normal form" of the metric $g = -dt^2 + \tilde{g}_{ij}(t,x)dx^i dx^j$, which $L = 1$ and $S = 0$ imply, see eq. (8.14) and Proposition 8.1 in §8.1.

[395] That is, the leaves of the canonical foliation around $\psi_t(\Sigma_0)$ are the images under ψ_t of the Σ_s in (8.253).

[396] Although $T_x\psi_t N_x$ may not equal $N_{\psi_t(x)}$, close to Σ it is still timelike. So if $Y_x \in T_x\Sigma_0$, so that $g_x(N_x, Y_x) = 0$, the vector $T_x\psi_t Y_x$ is orthogonal to some timelike vector and hence is spacelike, cf. Lemma 5.26 in O'Neill (1983).

Proof. Using Cartan's formula, the defining condition of a Gaussian vector field X becomes

$$0 = i_N \mathscr{L}_X g = (\mathscr{L}_X i_N + i_{[N,X]})g = (di_X + i_X d) i_N g + i_{[N,X]} g = d(i_X i_N g) + i_{[N,X]} g, \quad (8.257)$$

where $i_N g$ is the 1-form $g_{\mu\nu} N^\mu dx^\nu$, and we used $d i_N g = -ddt = 0$. Eq. (8.257) is the same as

$$[N,X] = -\nabla(g(N,X)), \quad (8.258)$$

which in turn may be rearranged as $\nabla_N S = \nabla_S N + \nabla L$. Taking normal and orthogonal components with respect to N and using (8.244), (8.254), as well as torsion-freeness, from which

$$[N,X] = \nabla_N X - \nabla_X N = \nabla_N X - \nabla_S N, \quad (8.259)$$

gives (8.256). These are first-order PDEs for L and S with given initial values $\tilde{L}$ and $\tilde{S}$ on Σ_0. In coordinates (t,x) adapted to the canonical foliation they even simplify to

$$\frac{\partial L}{\partial t} = 0; \qquad \frac{\partial S}{\partial t} = \tilde{\nabla} L. \quad (8.260)$$

The uniqueness claim then follows from the (elementary) theory of first-order PDEs. □

Corollary 8.13 *Let*

$$X_1 = L_1 N + S_1; \qquad X_2 = L_2 N + S_2 \quad (8.261)$$

be the unique Gaussian extensions of vector fields $\tilde{X}_1 = \tilde{L}_1 N + \tilde{S}_1$ and $\tilde{X}_2 = \tilde{L}_2 N + \tilde{S}_2$ defined on $\imath(\Sigma)$. Then, referring to (8.250), the commutator $[X_1, X_2]$ at $\imath(\Sigma)$ is given by (8.250), i.e.,

$$[X_1, X_2]_{|\imath(\Sigma)} = [(\tilde{L}_1, \tilde{S}_1), (\tilde{S}_1, \tilde{S}_2)]_\imath. \quad (8.262)$$

The proof is a simple computation, using (8.256). In (other) words, the curious bracket (8.250) is just the commutator of the Gaussian vector fields obtained by extending the given vector fields $\tilde{L}_1 N + \tilde{S}_1$ and $\tilde{L}_2 N + \tilde{S}_2$ on $\imath(\Sigma)$. Looking at a Gaussian vector field as an infinitesimal diffeomorphism of the special kind that preserves spacelike embeddings, this to some extent justifies calling (8.250) a "deformation algebra", although the situation remains to be clarified.

As already noted, the reason for studying this algebra is that in the Hamiltonian approach to GR the constraints reproduce it in the following sense: writing the total Hamiltonian $H \equiv H(\Sigma)$ as

$$H_{(\tilde{L},\tilde{S})}(\tilde{g},\tilde{\pi}) := \int_\Sigma \sqrt{\det(\tilde{g})}\,(L C_0(\tilde{g},\tilde{\pi}) + S^i C_i(\tilde{g},\tilde{\pi})), \quad (8.263)$$

see (8.218), (8.223) and (8.224), the canonical Poisson bracket will turn out to be

$$\{H_{(\tilde{L}_1,\tilde{S}_1)}, H_{(\tilde{L}_2,\tilde{S}_2)}\} = -H_{[(\tilde{L}_1,\tilde{S}_1),(\tilde{S}_1,\tilde{S}_2)]_\imath} = -H_{(\mathscr{L}_{\tilde{S}_1}\tilde{L}_2 - \mathscr{L}_{\tilde{S}_2}\tilde{L}_1, [\tilde{S}_1,\tilde{S}_2] + \tilde{L}_1\tilde{\nabla}\tilde{L}_2 - \tilde{L}_2\tilde{\nabla}\tilde{L}_1)}. \quad (8.264)$$

Writing $H_{\tilde{L}} := H_{(\tilde{L},0)}$ and $H_{\tilde{S}} := H_{(0,\tilde{S})}$, three interesting special cases of this bracket are

$$\{H_{\tilde{L}_1}, H_{\tilde{L}_2}\} = -H_{\tilde{L}_1\tilde{\nabla}\tilde{L}_2 - \tilde{L}_2\tilde{\nabla}\tilde{L}_1}; \quad (8.265)$$

$$\{H_{\tilde{S}_1}, H_{\tilde{S}_2}\} = -H_{[\tilde{S}_1,\tilde{S}_2]} = -H_{\mathscr{L}_{\tilde{S}_1}\tilde{S}_2}; \quad (8.266)$$

$$\{H_{\tilde{S}}, H_{\tilde{L}}\} = -H_{\mathscr{L}_{\tilde{S}}\tilde{L}}. \quad (8.267)$$

Of these, the first involves the metric and is generally seen as mysterious. In the next two sections we define a Poisson bracket for GR and try to explain the special status of the "super" Hamiltonian (8.263) in the light of the bracket (8.264) and the theory of the momentum map.[397]

[397] For another perspective see Głowacki (2021), who derives Definition 8.11 as a consistency condition between the $4d$ and $3+1$ descriptions of the dynamics that, unlike Theorem 8.2, also holds off the constraint surface.

8.9 Poisson brackets, constraints, and momentum map

The momentum map was originally introduced in the 1960s by Kostant and Souriau in the setting of symplectic geometry. It clarified especially the relationship between conserved quantities and symmetries, culminating in a Hamiltonian version of Noether's theorem.[398] The simplest setting for the momentum map, however, is in Poisson geometry, where the Poisson bracket is not seen as a concept derived from the symplectic structure, but stands on its own.[399]

Definition 8.14 *A* **Poisson bracket** *on a manifold P is a Lie bracket $\{-,-\}$ on the (real) vector space $C^\infty(P)$, such that for each $h \in C^\infty(P)$ the map*

$$X_h : f \mapsto \{h, f\} \tag{8.268}$$

defines a vector field on P, called the **Hamiltonian vector field** *of h. A manifold P equipped with a Poisson bracket is called a* **Poisson manifold**, *and $(C^\infty(P), \{\,,\,\})$ is a* **Poisson algebra**.

Unfolding, we have an bilinear map $\{-,-\} : C^\infty(P) \times C^\infty(P) \to C^\infty(P)$ that satisfies

$$\{g, f\} = -\{f, g\}; \tag{8.269}$$

$$\{f, \{g, h\}\} + \{h, \{f, g\}\} + \{g, \{h, f\}\} = 0; \tag{8.270}$$

$$\{f, gh\} = \{f, g\}h + g\{f, h\}, \tag{8.271}$$

where (8.269) - (8.270) is the Lie bracket property and (8.271) is the Leibniz rule for derivations.

The flow ψ_t of X_h is the motion generated by h, seen as "the Hamiltonian". Hence if (x^i) are coordinates on P, and we write $x(t)$ for $\psi_t(x)$, then $x(t)$ solves the coupled first-order ODEs

$$\frac{dx^i(t)}{dt} = \{h, x^i\}(x(t)). \tag{8.272}$$

The following result is crucial, although its proof is a straightforward exercise:

Proposition 8.15 *A Poisson bracket on P defines a Lie algebra homomorphism*

$$C^\infty(P) \to \mathfrak{X}(P); \qquad h \mapsto X_h. \tag{8.273}$$

In particular, for any $f, g \in C^\infty(P)$ we have

$$[X_f, X_g] = X_{\{f,g\}}. \tag{8.274}$$

The oldest example of a Poisson manifold is $P = \mathbb{R}^{2n}$ (even $n = 1$ is interesting!), where

$$\{f, g\} = \sum_{j=1}^{n} \left(\frac{\partial f}{\partial p_j}\frac{\partial g}{\partial q^j} - \frac{\partial f}{\partial q^j}\frac{\partial g}{\partial p_j} \right). \tag{8.275}$$

In that case, the Hamiltonian vector field of h is obviously given by

$$X_h = \sum_{j=1}^{n} \left(\frac{\partial h}{\partial p_j}\frac{\partial}{\partial q^j} - \frac{\partial h}{\partial q^j}\frac{\partial}{\partial p_j} \right). \tag{8.276}$$

[398] See Souriau (1969), Kostant (1970), Kijowski & Tulczyjew (1979), Guillemin & Sternberg (1982), Abraham & Marsden (1985), Marsden & Ratiu (1999), and Ortega & Ratiu (2004).

[399] In this approach symplectic geometry is a special case of Poisson geometry, arising when the Poisson tensor Π is invertible. The symplectic form ω is then the inverse of Π and the Poisson bracket equals $\{f, g\} = \omega(X_f, X_g\}$.

Writing $\psi_t(p,q) = (p(t),q(t))$, we see that this flow is given by ***Hamilton's equations***:

$$\frac{dp_j(t)}{dt} = \{h,p_j\}(p(t),q(t)) = -\frac{\partial h(p(t),q(t))}{\partial q^j}; \tag{8.277}$$

$$\frac{dq^j(t)}{dt} = \{h,q^j\}(p(t),q(t)) = \frac{\partial h(p(t),q(t))}{\partial p_j}. \tag{8.278}$$

A different kind of example is $P = \mathbb{R}^3$, which is *odd-dimensional*, where we define

$$\{f,g\}(x,y,z) = \\ x\left(\frac{\partial f}{\partial y}\frac{\partial g}{\partial z} - \frac{\partial f}{\partial z}\frac{\partial g}{\partial y}\right) + y\left(\frac{\partial f}{\partial z}\frac{\partial g}{\partial x} - \frac{\partial f}{\partial x}\frac{\partial g}{\partial z}\right) + z\left(\frac{\partial f}{\partial x}\frac{\partial g}{\partial y} - \frac{\partial f}{\partial y}\frac{\partial g}{\partial x}\right). \tag{8.279}$$

This is a special case of a general construction. Let $\mathfrak{g}$ be a Lie algebra, with basis (T_a), so that

$$[T_a,T_b] = \sum_c C_{ab}^c T_c, \tag{8.280}$$

for certain ***structure constants*** C_{ab}^c. We write θ in the dual vector space $\mathfrak{g}^*$ as $\theta = \sum_a \theta_a \omega^a$, where (ω_a) is the dual basis to a chosen basis of $\tilde{g}$, i.e., $\omega^a(T_b) = \delta_b^a$. In terms of these coordinates, the ***Lie–Poisson bracket*** on $C^\infty(\mathfrak{g}^*)$ is defined by the formula

$$\{f,g\}(\theta) = C_{ab}^c \theta_c \frac{\partial f(\theta)}{\partial \theta_a}\frac{\partial g(\theta)}{\partial \theta_b}. \tag{8.281}$$

Without a basis of $\mathfrak{g}$, the Lie–Poisson bracket may also be defined by extending the formula

$$\{\hat{A},\hat{B}\} = \widehat{[A,B]}, \tag{8.282}$$

where $A,B \in \mathfrak{g}$ and $\hat{A} \in C^\infty(\mathfrak{g}^*)$ is the evaluation map $\hat{A}(\theta) = \theta(A)$.

We now turn to the momentum map, which generalizes momentum, angular momentum, and almost every other quantity related to symmetry and conservation laws, culminating in Noether's Theorem. First, independently of Lie *groups*, Lie *algebras* also "act" on manifolds:

Definition 8.16 *Let $\mathfrak{g}$ be a Lie algebra and P a manifold. A* **$\mathfrak{g}$-action** *on P is a Lie algebra homomorphism from $\mathfrak{g}$ to $\mathfrak{X}(P)$, written $A \mapsto \xi_A$, so that in particular,*

$$[\xi_A,\xi_B] = \xi_{[A,B]}. \tag{8.283}$$

If $\mathfrak{g} = \mathrm{Lie}(G)$, then such actions usually arise from G-actions via (8.246), i.e. $\xi_A = \varphi_*(A)$.

Definition 8.17 *A* **momentum map** *for a Lie algebra action on a Poisson manifold P is a map*

$$J : P \to \mathfrak{g}^*, \tag{8.284}$$

such that, defining $J_A : P \to \mathbb{R}$ by $J_A(x) = \langle J(x),A\rangle \equiv J(x)(A)$, for each $A \in \mathfrak{g}$ we have

$$\xi_A = X_{J_A} := \{J_A,-\}. \tag{8.285}$$

A Lie algebra action with momentum map is called **Hamiltonian**.

In words, for any $A \in \mathfrak{g}$, taking the Poisson bracket with the function J_A generates the flow in P obtained by acting on P with the one-parameter subgroup $s \mapsto \exp(-sA)$ of G. ***Noether's (first) theorem*** then gives the familiar link between symmetries and conserved quantities:

Theorem 8.18 *Let P be a Poisson manifold with an action of a connected Lie group G, whose associated $\mathfrak{g}$-action (8.246) has a momentum map $J : P \to \mathfrak{g}^*$. If $h \in C^\infty(P)$ is G-invariant, i.e.*

$$h(\gamma \cdot x) = h(x) \tag{8.286}$$

for each $\gamma \in G$ and $x \in X$, then for each $A \in \mathfrak{g}$, the function J_A is constant along the flow ψ_t of X_h. That is, for any $x \in P$ and any $t \in \mathbb{R}$ for which $\psi_t(x)$ is defined, we have

$$J_A(\psi_t(x)) = J_A(x). \tag{8.287}$$

Proof. Using all assumptions, as well as the definition of a flow, we compute:

$$\begin{aligned}
\frac{d}{dt} J_A(\psi_t(x)) &= X_h J_A(\psi_t(x)) && (\psi_t \text{ is flow of } X_h) \\
&= \{h, J_A\}(\psi_t(x)) && (\text{definition of } X_h) \\
&= -\{J_A, h\}(\psi_t(x)) && (\text{antisymmetry of bracket}) \\
&= X_{J_A} h(\psi_t(x)) && (\text{definition of } X_{J_A}) \\
&= -\xi_A h(\psi_t(x)) && (8.285) \\
&= -\frac{d}{ds} h(e^{-sA} \psi_t(x))_{|s=0} && (8.246) \\
&= -\frac{d}{ds} h(\psi_t(x))_{|s=0} && G\text{-invariance of } h \\
&= 0. && \square
\end{aligned}$$

A simple example is

$$P = \mathbb{R}^6 = \mathbb{R}^3 \times \mathbb{R}^3, \tag{8.288}$$

with coordinates $x = (\vec{p}, \vec{q})$, where $\vec{p} = (p_1, p_2, p_2)$ and $\vec{q} = (q^1, q^2, q^3)$, equipped with the "canonical" Poisson bracket (8.275).

- Let $G = \mathbb{R}^6$ (as an additive group) act on P by

$$(\vec{a}, \vec{b}) \cdot (\vec{p}, \vec{q}) = (\vec{p} + \vec{a}, \vec{q} + \vec{b}). \tag{8.289}$$

  Then the derived $\mathfrak{g}$-action has a momentum map: identifying $\mathfrak{g} \cong \mathfrak{g}^* \cong \mathbb{R}^6$, this is

$$J(\vec{p}, \vec{q}) = (\vec{q}, -\vec{p}), \tag{8.290}$$

  and if the (sub)group $G = \mathbb{R}^3$ acts on P by

$$\vec{b} : (\vec{p}, \vec{q}) \mapsto (\vec{p}, \vec{q} + \vec{b}), \tag{8.291}$$

  we simply have

$$J(\vec{p}, \vec{q}) = -\vec{p}. \tag{8.292}$$

  The minus sign is of course unfortunate, but repairing this gives other undesirable signs elsewhere. By Noether's Theorem, if the potential V in a Hamiltonian

$$h(p, q) = p^2/2m + V(q) \tag{8.293}$$

is translation-invariant, then momentum $\vec{p}$ is conserved. Similarly, if $G = SO(3)$ acts on the same phase space $\mathbb{R}^6$ by

$$R \cdot (\vec{p}, \vec{q}) = (R\vec{p}, R\vec{q}), \tag{8.294}$$

then the derived $\mathfrak{g}$-action has a momentum map, which, identifying $\mathfrak{so}(3)^* \cong \mathbb{R}^3$, equals

$$J(\vec{p}, \vec{q}) = -\vec{q} \times \vec{p}, \tag{8.295}$$

which is (minus) the angular momentum! This time, if in the above Hamiltonian the potential V is rotation-invariant, Noether makes angular momentum $\vec{q} \times \vec{p}$ conserved.

- Now we keep $G = SO(3)$ but change P to $\mathbb{R}^3$ with the Poisson bracket (8.279) and take the defining action of $G = SO(3)$. If again we identify $\mathfrak{so}(3)^* \cong \mathbb{R}^3$, this action has a momentum map $J : \mathbb{R}^3 \to \mathbb{R}^3$, given by

$$J(\vec{x}) = \vec{x}. \tag{8.296}$$

More generally, the momentum map for the coadjoint action of G on $\mathfrak{g}^*$, with Poisson bracket (8.282), is simply the identity map $\mathfrak{g}^* \to \mathfrak{g}^*$, i.e.,

$$J(\theta) = \theta; \qquad \Leftrightarrow \qquad J_A = \hat{A}. \tag{8.297}$$

As a crucial point, it would be natural to expect that a momentum map J, if it exists, satisfies

$$\{J_A, J_B\} = J_{[A,B]} \tag{8.298}$$

for all $A, B \in \mathfrak{g}$. This property holds in our examples so far *except the first*,[400] even on $P = \mathbb{R}^2$: since $G = \mathbb{R}^2$ is abelian we have $[A, B] = 0$ and hence $J_{[A,B]} = 0$, but in a suitable basis (e_1, e_2) of $\mathfrak{g} = \mathbb{R}^2$ we have $J_{e_1} = q$ and $J_{e_2} = -p$, so that $\{J_{e_1}, J_{e_2}\} = 1$, i.e. the unit function on $\mathbb{R}^2$. However, one may always restore (8.298) by passing to a suitable central extension G (and $\mathfrak{g}$); in the case at hand this is the $(3d)$ Heisenberg group (see the references in footnote 398).

We now *try* to understand the Poisson bracket (8.264) of canonical GR in this light. We first *define* the bracket (and the underlying phase space) in question. Fixing Σ, we write $\mathscr{R} \equiv \mathscr{R}(\Sigma)$ for the space of smooth $3d$ Riemannian metrics on Σ. The associated tangent bundle is

$$T\mathscr{R} \cong \mathscr{R} \times \mathscr{S}_2 \subset \mathscr{S}_2 \times \mathscr{S}_2, \tag{8.299}$$

where $\mathscr{S}_2 = T^{(2,0)}(\Sigma)$ is the vector space of all covariant 2-tensors t_{ij} on Σ. Similarly, the associated cotangent bundle may be written in (cartesian) product form as

$$T^*\mathscr{R} \cong \mathscr{R} \times \mathscr{S}_d^2 \subset \mathscr{S}_2 \times \mathscr{S}_d^2, \tag{8.300}$$

where $\mathscr{S}_d^2 = T_d^{(0,2)}(\Sigma)$ is the vector space of all contravariant 2-densities d_{ij} on Σ.[401] Then

$$\langle t, d \rangle = \int_\Sigma t_{ij} d^{ij}, \tag{8.301}$$

[400] If G is connected and the given $\mathfrak{g}$-action on P comes from a G-action via (8.246), then (8.298) holds iff the G-action is equivariant with respect to the coadjoint action on $\mathfrak{g}^*$ (i.e. the dual to the adjoint action of G on $\mathfrak{g}$).

[401] Assuming Σ orientable, these are 3-form valued covariant 2-tensors, or equivalently tensor products of covariant 2-tensors and 3-forms, cf. §7.1, so that after contraction of the indices they can be integrated over Σ. In coordinates one may assume $dx^1 \wedge dx^2 \wedge dx^3$ as a standard volume form and hence $\mathscr{S}_d^2 \cong \mathscr{S}_d^2$ but not canonically so. Strictly speaking, in formulae like (8.209) one should then write $\sqrt{\tilde{g}} d^3x$ instead of $\sqrt{\tilde{g}}$.

where $t_{ij} \in T_{\tilde{g}}\mathscr{R}$ and $d^{ij} \in T^*_{\tilde{g}}\mathscr{R}$, defines a pairing between $\mathscr{S}_2$ and $\mathscr{S}^2_d$.[402] Writing elements of the cotangent bundle $T^*\mathscr{R}$ as $(\tilde{g},\tilde{\pi})$, we also have

$$T_{(\tilde{g},\tilde{\pi})}T^*\mathscr{R} \cong \mathscr{S}_2 \times \mathscr{S}^2_d. \tag{8.302}$$

We may turn $T^*\mathscr{R}$ into a Poisson manifold by generalizing the canonical bracket (8.275) to

$$\{f,g\} = \int_\Sigma \left(\frac{\delta f}{\delta \tilde{\pi}^{ij}} \frac{\delta g}{\delta \tilde{g}_{ij}} - \frac{\delta g}{\delta \tilde{\pi}^{ij}} \frac{\delta f}{\delta \tilde{g}_{ij}} \right), \tag{8.303}$$

where the δ are functional derivatives. This informal expression should be made precise. We only (need to) consider functions $f : T^*\mathscr{R} \to \mathbb{R}$ of the form $f = \int_\Sigma F$, in which

$$F : T^*\mathscr{R} \to C^\infty_d; \qquad C^\infty_d \equiv C^\infty_d(\Sigma) := C^\infty(\Sigma) \otimes \Lambda^3(\Sigma), \tag{8.304}$$

i.e. the space of density-valued smooth functions on Σ, so that the integral $\int_\Sigma F$ is defined. Using (8.300), and assuming (typically Sobolev norm) topologies on all spaces involved for taking limits, functions F defined as in (8.304) then have partial Fréchet derivatives

$$D_{\tilde{\pi}} f(\tilde{g},\tilde{\pi}) : \mathscr{S}^2_d \to C^\infty_d; \qquad D_{\tilde{g}} f(\tilde{g},\tilde{\pi}) : \mathscr{S}_2 \to C^\infty_d, \tag{8.305}$$

defined by

$$D_{\tilde{\pi}} f(\tilde{g},\tilde{\pi})(\rho) := \lim_{t\to 0} \frac{F(\tilde{g},\tilde{\pi}+t\rho) - F(\tilde{g},\tilde{\pi})}{t}; \tag{8.306}$$

$$D_{\tilde{g}} f(\tilde{g},\tilde{\pi})(h) := \lim_{t\to 0} \frac{F(\tilde{g}+th,\tilde{\pi}) - F(\tilde{g},\tilde{\pi})}{t}, \tag{8.307}$$

etc. Writing C^∞ for $C^\infty(\Sigma)$ as is customary in this business, these maps have (smooth) duals

$$D_{\tilde{\pi}} f(\tilde{g},\tilde{\pi})^* : C^\infty \to \mathscr{S}_2; \qquad D_{\tilde{g}} f(\tilde{g},\tilde{\pi})^* : C^\infty \to \mathscr{S}^2_d, \tag{8.308}$$

respectively, with respect to the natural L^2 pairing, cf. (8.301). Following Fischer and Marsden, the Poisson bracket of $f = \int_\Sigma F$ and $g = \int_\Sigma G$ on $T^*\mathscr{R}$ is then rigorously defined by

$$\{f,g\}(\tilde{g},\tilde{\pi}) := \int_\Sigma \langle D_{\tilde{\pi}} F(\tilde{g},\tilde{\pi})^*(1_\Sigma), D_{\tilde{g}} G(\tilde{g},\tilde{\pi})^*(1_\Sigma)\rangle. \tag{8.309}$$

With respect to this Poisson bracket, lengthy computations recover (8.265) - (8.267), where $(\tilde{L},\tilde{S})$ are the values of the lapse and shift (L,S) at any fixed time t and do not necessarily come from a Gaussian vector field. Furthermore, analogous computations bring the Hamiltonian equations of motion (8.227) - (8.228) in the Poisson bracket form given by (8.277) - (8.278), i.e.

$$\frac{\partial \tilde{g}_{ij}}{\partial t} = \{H_{(L,S)}, \tilde{g}_{ij}\}; \qquad \frac{\partial \tilde{\pi}^{ij}}{\partial t} = \{H_{(L,S)}, \tilde{\pi}^{ij}\}, \tag{8.310}$$

where we write $H_{(L,S)}$ instead of $H_{(\tilde{L},\tilde{S})}$ in order to make clear that the lapse and shift (L,S) are arbitrary (as long as they come from a regular foliation, if only a local one).[403] It is crucial that the equivalence between (8.227) - (8.228) and (8.310) holds "off-shell", i.e. whether or not the constraints are valid and hence whether or not the ensuing space-time metric g is Ricci-flat.

[402] Since we are at $\tilde{g} \in \mathscr{R}$ from which the standard integration with respect to $\sqrt{\tilde{g}}d^3x$ is defined, one may similarly pair $\mathscr{S}_2$ with the space $\mathscr{S}^2 = T^{(0,2)}(\Sigma)$ of "ordinary" contravariant 2-densities.

[403] These computations are done (though never fully) in Arnowitt, Deser, & Misner (1962), DeWitt (1967), Misner, Thorne, & Wheeler (1973), §21.6, Fischer & Marsden (1979), Poisson (2004), §4.2, and Thiemann (2007), §1.5.

8.10 A momentum map for canonical general relativity?

The combination of (8.225), (8.250), (8.298), and (8.310) makes it attractive to regard the Hamiltonian (8.263) as a momentum map of some kind. The point is not just that the various Lie and Poisson brackets match,[404] but also that the role of the lapse and the shift (L,S), which appear as parameters in the Hamiltonian, is now clearly distinguished from the role of $(\tilde{g},\tilde{\pi})$:

- The point $(\tilde{g},\tilde{\pi}) \in T^*\mathscr{R}$ is simply the *argument* x of $H = J_A(x)$;
- The lapse and shift (L,S) play the role of the *label* $A \in \mathfrak{g}$, cf. Definition 8.17.

The original idea of Fischer and Marsden to do so was as follows.[405] With Poisson manifold

$$P = T^*\mathscr{R}, \tag{8.311}$$

take $(\tilde{g}_0,\tilde{\pi}_0) \in P$, and some MGHD $(M,g,\imath)$ of the associated initial data $(\Sigma,\tilde{g}_0,\tilde{k}_0)$. We relabel $\imath$ as $\imath_0$ since it will act as a "reference embedding"; by definition $(M,g,\imath_0)$ induces the initial data $(\tilde{g}_0,\tilde{k}_0)$ on $\imath_0(\Sigma) \subset M$. Then $\imath \in \mathrm{Emb}(\Sigma,M,g)$ sends $(\tilde{g}_0,\tilde{\pi}_0)$ to the point $(\tilde{g},\tilde{\pi}) \in P$ obtained from the data induced by $(M,g,\imath)$ on $\imath(\Sigma)$. As we have seen, tangent vectors to curves in $\mathrm{Emb}(\Sigma,M,g)$ may be identified with pairs $(\tilde{L},\tilde{S}) \in C^\infty(\Sigma) \times \mathfrak{X}(\Sigma)$, and if we agree that these pairs form something like a Lie algebra $\mathfrak{g}$ of $\mathrm{Emb}(\Sigma,M,g)$, then these pairs will be labels A in J_A, as advertised above. Because of (8.310), the Hamiltonian is a momentum map for this Lie algebra, and because of (8.264) this momentum map even satisfies the pleasant relation (8.298).

Elegant as it is, this idea is questionable in (at least) two different ways:

1. The thing that acts, i.e. $\mathrm{Emb}(\Sigma,M,g)$, depends on the point $(\tilde{g}_0,\tilde{\pi}_0)$ of P at which the action is supposed to be defined.[406] To repair this, we will have to use (Lie) groupoids.
2. The MGHD $(M,g,\imath)$ is only defined when $(\tilde{g}_0,\tilde{\pi}_0)$ satisfies the constraints, and even so it is only defined up to isometry, see Theorem 7.10. Both problems can be addressed by refraining from the use of the MGHD, and even from only working with solutions to the (vacuum) Einstein equations. However, the constructions will then merely be local.[407]

We fix Σ and only consider space-times (M,g) that can be obtained from some $(\tilde{g},\tilde{\pi}) \in T^*\mathscr{R}$ by solving the coupled evolution equations (8.227) - (8.228) with lapse $L = 1$ and shift $S = 0$; this fixes some representative in the isometry class of (M,g). This can, in general, only be done locally in time, but since we will quickly pass to an infinitesimal level this is no problem; the entire Lie *groupoid* construction may merely be seen as motivation for the ensuing Lie *algebroid* construction. We may therefore assume that $M = I \times \Sigma$, where $0 \in I \subset \mathbb{R}$ is some open interval.

[404] The minus sign in (8.264) and hence the corresponding minus signs in (8.265) - (8.267) are caused by the fact that, as mentioned after (8.245), the Lie bracket in $\mathfrak{X}(M)$ seen as $\mathrm{Lie}(\mathrm{Diff}(M))$ is *minus* the commutator, and likewise for the Gaussian vector fields and for $\mathfrak{X}(\Sigma)$, so that (8.298) is correctly reproduced if we regard H as a momentum map J. Many authors, including Fischer & Marsden (1979), have the opposite sign for (8.268) as well as for the canonical Poisson bracket (8.275), in which case (8.263) - (8.267) has no minus signs, but (8.298) does.

[405] See Fischer & Marsden (1979), §4.6. The follow-up papers they referred to for details never appeared.

[406] This part of the construction might be justified in that the remaining steps, reviewed below, only depend on the orbits of the "action", rather than on the specific mathematical object that "acts" and causes these orbits.

[407] The ideas below are preliminary. For a different approach, so far also "work in progress" (though more advanced), see Blohmann, Barbosa Fernandes, & Weinstein (2013) and Blohmann & Weinstein (2018). For attacks on the problem based on (multi-)symplectic geometry rather than Lie groupoids see Kijowski & Tulczyjew (1979), Lee & Wald (1990), the legendary GIMMsy project (Gotay *et al.*, 1998–2004), and Forger & Romero (2005).

One obtains a solution to the vacuum Einstein equations in this way only if $(\tilde{g},\tilde{\pi})$ satisfies the constraints, but this is not necessary at this stage (and would even jeopardize the construction). Let $G_0 = \mathrm{Emb}(\Sigma)$ consist of all triples $(M,g,\imath)$, where (M,g) is some space-time of the said type and $\imath : \Sigma \hookrightarrow M$ is some spacelike embedding with respect to g. Let

$$G = \mathrm{Move}(\Sigma) := \mathrm{Emb}(\Sigma) \times \mathrm{Emb}(\Sigma) \tag{8.312}$$

be the associated pair groupoid:[408] elements $m \in G$ "move" some $\imath_1(\Sigma)$ to some $\imath_2(\Sigma)$. Let

$$p : T^*\mathscr{R} \to G_0 = \mathrm{Emb}(\Sigma); \qquad (\tilde{g},\tilde{\pi}) \mapsto (M,g,\imath) \tag{8.313}$$

be given by taking (M,g) to be the space-time obtained by solving the evolution equations (8.227) - (8.228) with lapse $L = 1$ and shift $S = 0$, and take $\imath(x) = (0,x)$. Then G acts on the map p in the natural way described above, i.e., the action of $((M_1,g_1,\imath_1),(M_2,g_2,\imath_2)) \in G$ on $(\tilde{g}_0,\tilde{\pi}_0) \in T^*\mathscr{R}$ is defined iff the triple $(M_2,g_2,\imath_2)$ equals $p(\tilde{g}_0,\tilde{\pi}_0) = (M,g,\imath)$ as just described, in which case, as in Fischer–Marsden, the result is the pair $(\tilde{g},\tilde{\pi})$ induced by g_1 on $\imath_1(\Sigma) \subset M_1$.

With an appropriate smooth or diffeological structure, the Lie algebroid $\pi : \mathfrak{g} \to G_0$ associated to G is the tangent bundle $T\mathrm{Emb}(\Sigma) \to \mathrm{Emb}(\Sigma)$, which, at fixed $(M,g,\imath)$, we have studied in some detail above. Consequently, the given G-action on $p : T^*\mathscr{R} \to \mathrm{Emb}(\Sigma)$ induces a $\mathfrak{g}$-action

$$\xi : \mathfrak{X}(\mathrm{Emb}(\Sigma)) \to \mathfrak{X}(T^*\mathscr{R}); \qquad A \mapsto \xi_A, \tag{8.314}$$

where, according to eq. (8.243) in §8.8, a vector field A on $\mathrm{Emb}(\Sigma)$ associates a Gaussian vector field (L,S), or equivalently its initial data $(\tilde{L},\tilde{S})$ at $\imath(\Sigma)$, to a triple $(M,g,\imath)$, cf. Proposition 8.12. Writing $X = LN + S$ as before, and letting tildes denote the restrictions of the given quantities to $\imath(\Sigma)$, the map ξ defining the $\mathfrak{g}$-action is then quite beautifully given by

$$\xi_A : (\tilde{g},\tilde{\pi}) \mapsto (\widetilde{\mathscr{L}_X g},\widetilde{\mathscr{L}_X \pi}); \qquad X = A(p(\tilde{g},\tilde{\pi})). \tag{8.315}$$

[408] A ***groupoid*** is a small category with inverses, i.e. one has two sets G and G_0 (which in our case are infinite-dimensional manifolds whose smooth or diffeological structure remains to be developed, cf. footnote 391), with maps $i : G_0 \to G$ (the *unit*), $s,t : G \to G_0$ (*source* and *target*), $\mu : G \times_{G_0} G \to G$ (*multiplication*), where $G \times_{G_0} G := \{(x,y) \in G \times G \mid s(x) = t(y)\}$, and $I : G \to G$ (*inverse*), such that, writing $xy := \mu(x,y)$ whenever defined, we have $s(xy) = s(y)$, $t(xy) = t(x)$, $t \circ I = s$, $s \circ I = t$, $(xy)z = x(yz)$, $s \circ i = t \circ i = \mathrm{id}_{G_0}$, $xi(s(x)) = i(t(x))x = x$, $I(x)x = i(s(x))$, and $xI(x) = i(t(x))$. Thus G consists of arrows x sending $s(x) \in G_0$ to $t(x) \in G_0$. For example, each equivalence relation $\sim$ on some set G_0, i.e. each subset $R \subset G_0 \times G_0$, defines a groupoid $G = R$ with structure borrowed from the simplest example $R = G_0 \times G_0$, called the ***pair groupoid*** on G_0, where $s(a,b) = b$, $t(a,b) = a$, $i(a) = (a,a)$, $I(a,b) = (b,a)$, and $(a,b) \cdot (b,c) = (a,c)$. The "opposite" example is a group, where $G_0 = \{e\}$. A groupoid G on G_0 may act not so much on a space but on a *map* $p : P \to G_0$, via a map $\varphi : G \times_{G_0} P \to P$, where $G \times_{G_0} P := \{(x,\rho) \in G \times P \mid s(x) = p(\rho)\}$, subject to $t(x\rho) = p(\rho)$, where we write $x\rho = \varphi(x,\rho)$, $(xy)\rho = x(y\rho)$, and $i(a)\rho = \rho$, whenever defined. This is in fact the key to the use of groupoids in our GR context, since we see that, except when $G_0 = \{e\}$, only part of G acts on a given point $p \in P$ and this part may very well depend on p. In the presence of sufficient smoothness a groupoid–then called a ***Lie groupoid***–has an associated ***Lie algebroid*** $\pi : \mathfrak{g} \to G_0$, which is a vector bundle over G_0, equipped with an additional map $\alpha : \mathfrak{g} \to TG_0$ (called the ***anchor***) and a Lie bracket on $\Gamma(\mathfrak{g})$, the space of smooth sections of π, such that $[A,fB] = f[A,B] + \alpha(A)f \cdot B$ for each $f \in C^\infty(G_0)$, and $\alpha([A,B]) = [\alpha(A),\alpha(B)]$. Here the simplest examples are $\pi : TG_0 \to G_0$ with trivial anchor, which arises as the Lie algebroid of the pair groupoid on G_0, and the Lie algebra of a Lie group, seen here as a vector "bundle" on a point (and hence as a vector space). As in the Lie group case, one has a notion of a $\mathfrak{g}$-action on a map $p : P \to G_0$, which often comes from a G-action but is defined independently of such an origin. Thus we have a Lie algebra homomorphism $\xi : \Gamma(\mathfrak{g}) \to \mathfrak{X}(P)$, $A \mapsto \xi_A$, such that $\xi_{fA} = (p^*f)\xi_A$ and $\xi_A(\pi^*f) = \alpha(A)f$.

See Mackenzie (2005) for a comprehensive treatment of Lie groupoids, Lie algebroids, and their actions. Moerdijk & Mrcun (2003) and perhaps Landsman (1998, §III.3) or (2017, §7.4, §C.16) provide concise introductions.

The equivalence between (8.227) - (8.228) and (8.310) implies that this $\mathfrak{g}$-action has our Hamiltonian H as a momentum map.[409] For $X = S$ we have $\widetilde{\mathscr{L}_S g} = \mathscr{L}_{\tilde{S}}\tilde{g}$ and $\widetilde{\mathscr{L}_S \pi} = \mathscr{L}_{\tilde{S}}\tilde{\pi}$, so that ξ degenerates to a map $\tilde{\xi} : \mathfrak{X}(\Sigma) \to \mathfrak{X}(T^*\mathscr{R})$, $\tilde{\xi}_{\tilde{S}} : (\tilde{g},\tilde{\pi}) \mapsto (\mathscr{L}_{\tilde{S}}\tilde{g}, \mathscr{L}_{\tilde{S}}\tilde{\pi})$, which is the map obtained from (8.246) by taking $G = \mathrm{Diff}(\Sigma)$, acting on $N = T^*\mathscr{R}$ by pullback of its action $\varphi(\tilde{g}) = (\varphi^{-1})^*\tilde{g}$ on $\mathscr{R}$. With $\tilde{g} = \mathfrak{X}(\Sigma)$, it follows that $\tilde{S} \mapsto H_{0,\tilde{S}}$ is a momentum map for $\tilde{\xi}$.

In the general case we are back to Fischer and Marsden, who–and this may even have been their main goal–now invoke a powerful construction due to Marsden and Weinstein, namely ***symplectic reduction***.[410] In its simplest version, a Lie group G acts on a symplectic manifold P (i.e. a Poisson manifold whose Poisson tensor is invertible) with momentum map $J : P \to \mathfrak{g}^*$ satisfying (8.298). On suitable regularity assumptions,[411] the ***symplectic quotient***

$$P//G := J^{-1}(0)/G \tag{8.316}$$

has a unique invertible Poisson tensor $\tilde{\Pi}$ whose associated symplectic form $\tilde{\omega} = \Pi^{-1}$ satisfies

$$\pi^*_{J^{-1}(0)\to P//G}\tilde{\omega} = i^*_{J^{-1}(0)\hookrightarrow P}\omega, \tag{8.317}$$

where Π is the given invertible Poisson tensor on P with inverse $\omega = \Pi^{-1}$, and the notation is hopefully self-evident.[412] Under the stated regularity assumptions, $P//G$ is a (symplectic) manifold, which in case G defines gauge symmetries is identified with the space of physical degrees of freedom.[413] Furthermore, at each $x \in J^{-1}(0)$ the tangent space T_xP decomposes as

$$T_xP = T_x(J^{-1}(0)) \oplus T_xR = T_x(P//G) \oplus T_x(Gx) \oplus T_xR, \tag{8.318}$$

where T_xR is any (linear) complement of $T_x(J^{-1}(0))$ within T_xP (if one has a positive definite metric on T_xP, one may define such complements as orthogonal complements). Furthermore, since $T_x(Gx)$, i.e. the tangent space to the orbit through x, is a subspace of $T_x(J^{-1}(0))$, the latter splits into $T_x(Gx)$ and a complement thereof, which we may identify with $T_x(P//G)$.

Apart from problems arising from the non-validity of some of the technical assumptions that underwrite it (including the lack of a sufficiently developed smooth or diffeological framework so far), the above constructions, originally intended for finite-dimensional Lie groups G acting on finite-dimensional manifolds P, at least conceptually generalize to the infinite-dimensional phase space $P = T^*\mathscr{R}$ and our infinite-dimensional (Lie) groupoid $G = \mathrm{Move}(\Sigma, M)$. For $x = (\tilde{g},\tilde{\pi})$ lying in the constraint surface $H = 0$ (in which case, we recall, the ensuing space-time (M,g) solves the vacuum Einstein equations), the orbit $G\cdot(\tilde{g},\tilde{\pi})$ by construction consists of all initial data obtained from all spacelike embeddings $\iota : \Sigma \hookrightarrow M$ for the given metric g (i.e. on M).

---

[409] This is meant in the simplest way here, as a map $J : T^*\mathscr{R} \to \mathfrak{g}^*$. In the context of Lie groupoid actions there are various other, more refined notions of momentum maps, see e.g. Bos (2007) and Blohmann & Weinstein (2018).

[410] The original sources are Meyer (1973) and Marsden & Weinstein (1974). For a historical survey of symplectic reduction see Marsden & Weinstein (2001). See also the references in footnote 398, as well as Landsman (1998).

[411] These are that $0 \in \mathfrak{g}^*$ is a regular value of J and that G acts freely and properly at least on $J^{-1}(0)$.

[412] Poisson purists will find it preferable to write this construction in terms of the Poisson structure alone, but this only seems possible by appealing to the symplectic stratification theorem for Poisson manifolds, which also involves symplectic geometry. See e.g. Ortega & Ratiu (2004), §10.1. First, there is a unique Poisson bracket on P/G such that $\pi^*_{P\to P/G}\{f,g\}_{P/G} = \{\pi^*_{P\to P/G}f, \pi^*_{P\to P/G}g\}_P$. Second, $J^{-1}(0)/G$ is one of the symplectic leaves in P/G, with its associated Poisson structure (which by construction is symplectic) inherited from the one on P/G.

[413] This is more or less the definition of a gauge symmetry! Even if G gives observable changes, the quotient $P//G$ is useful for simplifying the equations of motion, provided these come from a G-invariant Hamiltonian h on P via (8.272), since there exists a unique Hamiltonian $\tilde{h}$ on $P//G$ such that, just like (8.317), $\pi^*_{J^{-1}(0)\to P//G}\tilde{h} = i^*_{J^{-1}(0)\hookrightarrow P}h$, from which the motion on P may be (re)constructed (in case of gauge symmetry, $\tilde{h}$ is the physical Hamiltonian).

Deformations of the initial data $(\tilde{g},\tilde{\pi})$ tangent to the G-orbit in (8.318), i.e. along the subspace

$$T_x(Gx) = T_{(\tilde{g},\tilde{\pi})}(G\cdot(\tilde{g},\tilde{\pi})), \tag{8.319}$$

therefore give rise to the same space-time (M,g), at least up to isometry.[414] Even though $(\tilde{g},\tilde{\pi})$ satisfies the constraints, as we assume, and hence lies in $H^{-1}(0)$, the term T_xR in (8.318) consists of deformations of $(\tilde{g},\tilde{\pi})$ off the constraint surface $H=0$. These deformations lead to space-times not satisfying the Einstein equations and can be ignored. Finally, the term $T_{(\tilde{g},\tilde{\pi})}(T^*\mathscr{R}//G)$ gives the direction of deformations of $(\tilde{g},\tilde{\pi})$ that lie within the constraint surface. These give rise to space-times that satisfy the vacuum Einstein equations but are non-isometric to the (isometric) space-time(s) with initial data $(\Sigma,\tilde{g},\tilde{\pi})$. If all this can be made to work globally, which of course is a big "if", the "space of gravitational degrees of freedom" may then be identified with $\mathscr{C}/G$, where $\mathscr{C}\subset T^*\mathscr{R}$ is the constraint set $H=0$. Furthermore, if $\mathrm{Einstein}(M)$ is the space of all metrics g on M that arise as the MGHD of initial data $(\Sigma,\tilde{g},\tilde{\pi})$ satisfying the constraints (and hence the vacuum Einstein equations), at least for $M=\mathbb{R}\times\Sigma$ we should also have

$$\mathrm{Einstein}(M)/\mathrm{Diff}(M) \cong \mathscr{C}/G. \tag{8.320}$$

In defense of this canonical dream,[415] let us note that at least "the count is right": at each $x\in\Sigma$ we *a priori* have 12 degrees of freedom (d.o.f.), since both $\tilde{g}_{ij}$ and $\tilde{\pi}^{ij}$ (or, equivalently, $\tilde{k}_{ij}$) are symmetric 3×3 matrices, having 6 independent components each. Four constraints reduce this number to $12-4=8$, and four components of $X=(L,S)$ further reduce this to $8-4=4$, that is, the gravitational field has 2 *physical* d.o.f. per point (plus 2 associated momenta). This result had previously been derived in §8.5 on the basis of a linear approximation to the Einstein equations, which leads to the identification of these d.o.f. with the two helicity states of a massless helicity-2 particle (i.e. the graviton). Instead, the approach here is geometric and non-perturbative.[416]

Jerry Marsden (1942–2010) discussing the momentum map for GR *with the author in 1999.*

[414] It is in this sense that H is said to generate gauge transformations in GR. But it does not follow that moving initial data $(\Sigma,\tilde{g},\tilde{\pi})$ in M is unobservable or otherwise unphysical! See §8.11 for further discussion.

[415] See Fischer & Marsden (1979), p. 207. The left-hand side was taken up by Fischer & Moncrief (1996, 1997).

[416] The fact that the count of the d.o.f. of the gravitational field can be done in these two very different ways reflects a deep schism in the world of quantum gravity (Armas, 2021). The majority goes for string theory (a perturbative particle-physics based ideology), whereas a sizable minority prefers a non-perturbative geometric approach. This schism was already implicit in the almost simultaneous publication of the three masterpieces on GR by Weinberg (1972) on the one hand, and Misner, Thorne, & Wheeler (1973) and Hawking & Ellis (1973) on the other.

8.11 Epilogue: The problem of time

Although the previous analysis is preliminary and non-rigorous, we expect that its conclusion is independent of the details, in the sense that any satisfactory analysis of the (gauge and non-gauge) degrees of freedom of GR should lead to the same general picture (perhaps this is what we mean by 'satisfactory'). This enables us to put in our tuppence worth on the scarlet ***problem of time***.

The philosophical analysis of time is as old as philosophy itself, traditionally starting around 500 BC with the opposition between Heraclitus, who famously maintained that everything constantly changed, and Parmenides, who felt that if not change, then at least time was an illusion.[417] Jumping to the twentieth century, in what follows we focus on the implications:[418]

$$\textit{time} \quad \Rightarrow \quad \textit{change} \quad \Rightarrow \quad \textit{A-series} \quad \Rightarrow \quad \textit{B-series}, \qquad (*)$$

where we use the standard terminology in the philosophy of time, introduced by McTaggart:

- In the ***A-series***, events are ordered in a time series that goes from the past to the present and moves on towards the future. This ordering assumes the existence of a "moving now" and as such describes what has been called ***manifest time***, which is what we actually experience. With respect to the "now", any event lies either in the past, or in the present, or in the future, and this status changes as time flows, that is, as the "now" moves on.

- The ***B-series***, on the other hand, merely orders events according to their relative position, which can be either that they are simultaneous, or that one is earlier or later than the other.

In particular, there is no "now" in the B-series. One version of the "problem of time", then, is the claim that modern physics gives us a B-series, whereas everyday experience gives us an A-series. In other words, physics fails to incorporate the "now" that dominates our perception of time.[419] However, physics is still supposed to be *compatible* with an A-series, whose existence is merely foreign to its language. This problem is soft compared to the radical claim we will now discuss:

$$\textit{general relativity does not even provide a B-series} \qquad (!)$$

[417] This Pre-Socratic opposition between "becoming" and "being", or "change" and "existence", continued with Aristotle. This had disastrous consequences for mathematical physics. In his *Metaphysics*, Aristotle organized knowledge into something like a 2×2 matrix, where the axes are "changing/permanent" and "dependent/independent" (that is, of man). He put physics in the change & independent entry, whereas mathematics was supposed to be permanent & independent (the latter against Plato). See e.g. Gaukroger (2020). This classification held back the interaction between physics and mathematics for 2000 years, until initially Kepler and Galilei and subsequently Huygens and especially Newton recombined them and thus provided the basis for modern science.

[418] These implication were all proposed by McTaggart (1908, 1927). See also Dainton (2010). The only implication that really counts for our technical discussion is "time $\Rightarrow$ B-series", or rather its contrapositive "no B-series $\Rightarrow$ no time", but the chain in (*) is convenient in order to frame the overall problem of time. The first implication goes back at least to Aristotle (*Physics*, Book IV, chapter 11), see Shoemaker (1969) for a nice philosophical analysis. It would be denied by Newton (Rynasiewicz, 2014), but GR can deny it, too, as it admits static solutions (see §8.4). The point, however, is that according to the arguments reviewed and critiqued below GR admits no flow of time whether or not time requires change. Similarly, the second implication needs to be argued for, as McTaggart does at some length, but his target is the A-series, whose alleged incoherence allows him to disprove the existence of time. Instead, the argument in our main text concerns the B-series. It is remarkable that of the two great twentieth-century philosophical treatises about existence and time, both of which are hard-core specimens of "armchair" philosophy based on pure speculation, McTaggart (1921, 1927) has been very influential on discussions that *are* informed by modern science, whereas Heidegger (1927) has, rightly, been completely sidelined in the philosophy of science.

[419] This is the version of the problem addressed in Callender (2017), whose opening sentences deserve to be quoted: *'Time is a big invisible thing that will kill you. For that reason alone, one might be curious about what it is.'*

In particular, the causal picture of space-time, based on the relations I or J, i.e. on the partial orderings $x \ll y$ or $x \leqslant y$ (cf. §5.3), is a hallucination (or, in more diplomatic parlance, it is part of the "manifest" image of GR, as opposed to its "scientific" image). In reality, or so it is claimed, there is just a "frozen" initial data set $(\Sigma, \tilde{g}, \tilde{k})$ whose development into a space-time (M, g) is unphysical. In other words, despite the fact that relative to some foliation $M = \sqcup\Sigma_t$ (cf. chapter 8) the initial data $(\Sigma, \tilde{g}, \tilde{k})$ appear to move into new and usually different data $(\Sigma_t, \tilde{g}_t, \tilde{k}_t)$, these data are merely different descriptions of the same physical situation. If true, there is no change, and hence, taking the contrapositive of (*), no time either. Was Parmenides right, after all?

In the literature one finds the following two arguments for the timelessness of GR:[420]

- *Diffeomorphisms:* Since the Einstein equations of GR are invariant under diffeomorphisms, even for given initial data its solutions are unique only up to diffeomorphisms. Therefore, in order to save determinism, the "observables" of the theory must be diffeomorphism-invariant, too. This excludes any explicit time-dependence of physical quantities.

- *Constrained Hamiltonian dynamics:* In the Dirac–Bergmann approach to GR as a constrained Hamiltonian system time evolution is generated by the Hamiltonian constraint, which according to his formalism generates gauge transformations. Once again, in the interest of saving determinism the effect of such transformations is deemed unphysical.

The second argument is a "Hamiltonian shadow" of the first.[421] Both arguments are based on an interpretational move within a certain formalism that is not in fact implied by that formalism.

In the Hamiltonian approach this move–which we question–interprets different canonical data on the same gauge orbit as physically indistinguishable. Although this is indeed the case in electrodynamics, in GR the situation is quite different. The real sense in which moving from the canonical data $(\tilde{g}_0, \tilde{\pi}_0)$ on Σ to time-evolved data $(\tilde{g}(t), \tilde{\pi}(t))$ on the same Σ is a gauge transformation, is that both data sets give rise to–i.e. are initial data for–the same space-time (M, g). To clarify this matter, for the convenience of the reader we now rephrase Theorem 8.2, which may be seen as a corollary and reinterpretation of Theorem 7.10, in Hamiltonian form.

[420] In connection with relativity, the philosophical analysis of time goes back to the special theory, see e.g. Cassirer (1921), Bergson (1922), Schlick (1922), Reichenbach (1928), of whom the latter two also involve the general theory. See e.g. Ryckman (2018) and Stuur (2019) for recent historical and philosophical analysis. The problem of time in GR as discussed here has its historical roots in Einstein's hole argument (see §1.5) and the ensuing issue of general covariance (§1.10), but may more specifically be traced back to Bergmann (1958, 1961). From theoretical physics we cite the reviews by Isham (1992) and Kuchar (1992) as well as the monograph by Anderson (2017); see also Thiemann (2007). Defendants of claim (!) include Barbour (1999), Earman (2002), and Rovelli (2004). From the philosophical literature our views are closest to Butterfield (1984) and Healey (2002, 2004). See also Norton (2010), Pitts (2014), Gryb & Thébault (2016), Rovelli (2019), and Thébault (2021). Maudlin (2002) dismisses the Hamiltonian version of the problem of time in GR as '*completely phony*', but this verdict is predicated on his erroneous claim, made even twice, that the initial value problem of GR 'admits of a unique maximal solution'. See Theorem 7.10, whose *lack of absolute uniqueness* (replaced by uniqueness up to isometry) is nothing but Hilbert's Cauchy-problem version of the hole argument and hence may be seen as the root of the problem of time in the PDE approach to GR. Indeed, what makes the problem of time look genuine (though solvable) is that it pops up in almost any formulation of GR. However, Maudlin's discussion of the "diffeomorphism" version is actually quite good.

[421] One also sometimes finds a mixture of these arguments to the effect that the (formal) Hamiltonian H in GR generates (space-time) diffeomorphisms, but this is hard to make sense of. If H, taken to be (8.263), acts on phase space $T^*\mathscr{R}$ as defined in §8.9, then there simply is no notion of $4d$ diffeomorphisms. If what is meant is a rewriting of Theorem 8.2 in Hamiltonian form via (8.310), then, as explained in the main text, the lapse and shift have given values, and diffeomorphism invariance of the theory is broken by the ensuing foliation of space-time. In that setting, what remains of the idea that time evolution in GR is a diffeomorphism is that $\partial_t \tilde{g} = \mathscr{L}_{\partial_t} \tilde{g}$ is a Lie derivative–indeed an infinitesimal diffeomorphism!–, which is true for almost any quantity in almost any geometric theory.

Theorem 8.19 *Let (M,g) be a globally hyperbolic space-time equipped with a foliation (8.1) by spacelike Cauchy surfaces Σ_t, and associated lapse L and shift S. Let $(\tilde{g}(t),\tilde{\pi}(t))$ be the canonical data on Σ_t induced by the 4-metric g on M via the–equivalent–data $(\tilde{g}(t),\tilde{k}(t))$ consisting of the 3-metric and exterior curvature on Σ_t induced by g, cf. (8.209) - (8.210).*

- *Then g is a solution of the vacuum Einstein equations iff, cf. (8.223) - (8.224):*
  1. *For **some** t the pair $(\tilde{g}(t),\tilde{\pi}(t))$ satisfies the constraint equations $C_0 = 0$ and $C_i = 0$;*
  2. *The maps $t \mapsto \tilde{g}_{ij}(t)$ and $t \mapsto \tilde{\pi}_{ij}(t)$ satisfy the evolution equations (8.310), where the Hamiltonian $H_{(L,S)}$, indexed by the lapse L and shift S, is given by (8.263).*
- *Conversely, given canonical data $(\tilde{g},\tilde{\pi})$ on Σ satisfying the constraints, the evolution equations (8.310) with specified lapse L and shift S have a solution $(t \mapsto \tilde{g}_{ij}(t), t \mapsto \tilde{\pi}_{ij}(t))$, which is unique in its time domain and defines a globally hyperbolic space-time (M,g) with associated foliation that returns the solution as just described: for each t the pair $(\tilde{g}(t),\tilde{\pi}(t))$, though originally defined on Σ, are the initial data induced on Σ_t by g.*

In the context of Theorem 7.10 this space-time was only given up to initial-data-preserving isometries (as in Hilbert's version of the hole argument), but in the context of Theorem 8.19 this lack of uniqueness is avoided by an explicit choice of the lapse L and shift S. Here it is crucial to realize that in the Hamiltonian formalism (as presented in sections 8.7 to 8.10) the Hamiltonian generating the dynamics (which allegedly consists of unphysical gauge transformations) is not the Hamiltonian constraint (7.148) itself, but the function $H_{(L,S)}$ appearing in Theorem 8.19, which is indexed by a specific choice of the lapse L and shift S. As explained in §8.1, such a choice amounts to a foliation (8.1) of the space-time that the initial data $(\Sigma,\tilde{g}_t,\tilde{k}_t)$ give rise to.

This foliation is arbitrary (as long as its leaves are spacelike), but once it has been chosen (if only implicitly, via the lapse and shift), it sets the standard (or reference frame) against which time and change in time are measured.[422] The changes we observe in the context of GR, from the motion of the perihelion of Mercury to the expansion of the universe, are real, but their quantification is somewhat arbitrary in that it may depend on the foliation. In other words, *numerical indicators* of change may depend on the reference frame against which they are measured, but this is nothing new. The only–but of course crucial–difference with special relativity is that in *general* relativity the "now" has become even more flexible. In Newtonian space-time hyperplanes of simultaneity must be horizontal. In Minkowski space-time they are no longer unique and may be tilted. In GR, all sorts of curved hypersurfaces Σ_t are allowed: as we argued in §1.10, *this* is what makes GR general. But even within this increased arbitrariness, the causal structure of a space-time (M,g) is well defined and hence it is perfectly clear what moving forward in time means, namely moving along a future-directed timelike curve.

The dissection of the "diffeomorphism" argument against time is similar, to the effect that once again time and change are perfectly well defined in GR, but are quantified relative to a foliation or reference frame, and hence are less absolute than in pre-relativistic physics.[423]

In conclusion, we have argued that (compared to other theories) GR has no new features that should affect the philosophical analysis of time. It surely supports the B-series, and seems neutral about the existence of the A-series, i.e., about the reality of the "moving present".

[422] Einstein (1917a) and Hilbert (1917) held this view. Einstein imagined a 'reference mollusk' (*Bezugsmolluske*, *ibid.*, p. 67), whereas Hilbert prosaically realized the frame as a system of measuring rods and 'light-clocks'.

[423] See Maudlin (2002) and Healey (2004).

9 Black holes I: Exact solutions

The theory of black holes is an interplay between abstract arguments, like Penrose's singularity theorem, and concrete examples. This chapter is devoted to the latter (but the next one returns to the abstract theory). As a warm-up, we start with a simple example that has no true singularity but illustrates the remarkable interplay between coordinate singularities and horizons.[424]

9.1 De Sitter space revisited

Recall from (4.92) in §4.4 that the two-dimensional de Sitter space dS_1^2 (with unit radius $\rho = 1$) is defined as the surface $-x_0^2 + x_1^2 + x_2^2 = 1$ in $\mathbb{R}^3$ with metric inherited from the Minkowski metric $\eta = -dx_0^2 + dx_1^2 + dx_2^2$. It is a Lorentzian manifold with constant curvature $k = 1$, topologically $dS_1^2 \cong \mathbb{R} \times S^1$, cf. (4.94). Each of the following coordinatizations is useful:[425]

$$x_0 = \sinh\tau; \qquad x_1 = \cosh\tau\cos\psi; \qquad x_2 = \cosh\tau\sin\psi; \tag{9.1}$$

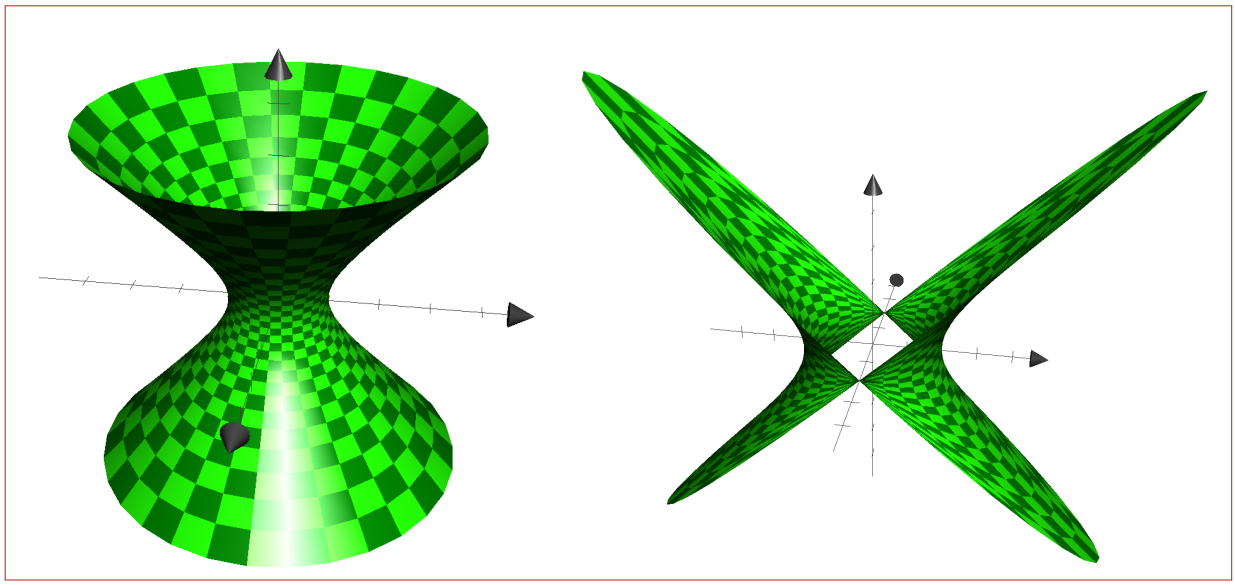
$$x_0 = \sinh t\cos\varphi; \qquad x_1 = \cosh t\cos\varphi; \qquad x_2 = \sin\varphi; \tag{9.2}$$

$$x_0 = \sinh t\sqrt{1-r^2}; \qquad x_1 = \cosh t\sqrt{1-r^2}; \qquad x_2 = r. \tag{9.3}$$

where $\tau, t \in \mathbb{R}$, ψ, φ in $(-\pi, \pi)$, and $r \in (-1, 1)$. In these coordinates, the de Sitter metric is

$$g_{dS} = -d\tau^2 + \cosh^2\tau\, d\psi^2 = -\sin^2\varphi\, dt^2 + d\varphi^2 = -f(r)dt^2 + f(r)^{-1}dr^2, \tag{9.4}$$

where $f(r) := 1 - r^2 = (1+r)(1-r)$, see also (6.2) and ensuing discussion.



*Two-dimensional de Sitter space embedded in three-dimensional Minkowski space. The left picture, extended to $\pm\infty$ along the x_0-axis (= z-axis), gives the complete space, as coordinatized by (9.1). The picture on the right (*idem dito*) is the part coordinatized by (9.2). It does* not *contain the "singular" points $(0,0,\pm1)$, so that the boundaries at $r = \pm1$ or $\varphi = \pm\frac{1}{2}\pi$ do not touch. Its right-hand part, called the* **static patch***, is the part coordinatized by (9.3), with a metric that at least looks singular at $r = \pm1$.*

[424] This section, which may be skipped at the expense of a cold start in §9.2, was inspired by §2 of Carter (1973). Carter discusses *anti* de Sitter space, technically even in a quite different way, but the spirit is similar.

[425] The first system can be extended to $\psi \in \mathbb{R}$, giving the metric on the universal covering $\widetilde{dS_1^2}$. We will not do so.

The first coordinate system (τ,ψ) covers the entire space, but it obscures the static nature of the metric. Staticity is obvious in the second system (t,φ), in which the timelike Killing vector field is given by ∂_t. The third system, which is obtained from the second by putting $\rho=\sin\varphi$ and restricting the range of φ to $(-\frac{1}{2}\pi,\frac{1}{2}\pi)$, is pedagogically useful as it gets us closer to the black hole solutions below. But it is also physically motivated, because the special values $r=\pm1$ and hence the boundaries of the region covered by the (r,t) coordinates correspond to a so-called ***Killing horizon***, where ∂_t becomes lightlike (see §10.8). Although this is suggested by the metric (9.4) we need better coordinates to establish this fact, since the horizon is not within the scope of the (r,t) system. To this end, anticipating the Schwarzschild black hole case, we solve

$$\frac{dr_*(r)}{dr}=\frac{1}{f(r)};\qquad\qquad r_*=\operatorname{arctanh} r=\tfrac{1}{2}\ln\left|\frac{1+r}{1-r}\right|, \tag{9.5}$$

where $r\in(-1,1)$ corresponds to $r_*\in\mathbb{R}$, with $r\to\pm1$ iff $r_*\to\pm\infty$. We proceed by introducing the analogue of the lightlike coordinates $u=t-r$ and $v=t+r$ in Minkowski space-time, i.e.

$$u=t-r_*,\qquad\qquad t=\tfrac{1}{2}(v+u); \tag{9.6}$$

$$v=t+r_*,\qquad\qquad r_*=\tfrac{1}{2}(v-u), \tag{9.7}$$

In terms of these, via the relation $r=\tanh r_*$ the metric is easily found to be

$$g_{dS}=-f(r)dudv=(1-\tanh^2(\tfrac{1}{2}(v-u)))dudv. \tag{9.8}$$

This expression is still singular as $r\to\pm1$. To remedy this at least near $r=+1$, we introduce

$$-U=e^u;\qquad\qquad V=e^{-v}, \tag{9.9}$$

which clearly satisfy $U<0$, $V>0$, and, in view of (9.9), (9.6) - (9.7), and (9.5), we have

$$UV=\exp(-2r_*)=\frac{r-1}{r+1}. \tag{9.10}$$

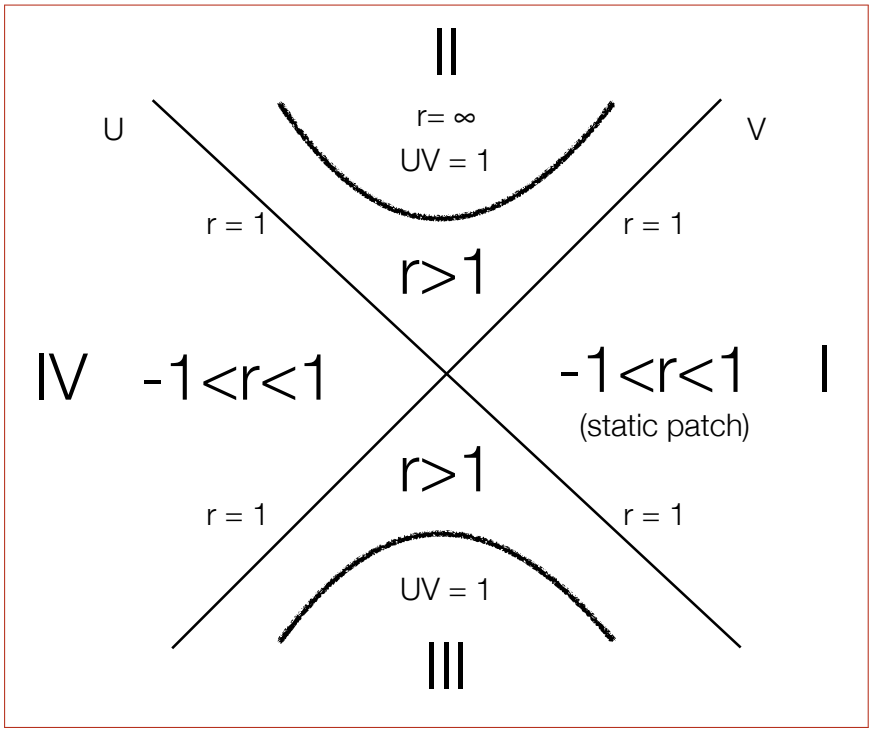
In terms of these coordinates, the metric is

$$g_{dS}=-\frac{4}{(1-UV)^2}dUdV, \tag{9.11}$$

and the coordinates (x_0,x_1,x_2) in terms of which dS_1^2 was originally defined are

$$x_0=\frac{-U-V}{1-UV};\qquad\qquad x_1=\frac{-U+V}{1-UV};\qquad\qquad x_2=\frac{1+UV}{1-UV}. \tag{9.12}$$

Since $r\to1$ corresponds to $r_*\to\infty$ and hence to $UV\to0$, the metric is now regular for $r\to1$. We can even pass through this barrier by allowing U to also be zero or positive, at least provided $UV<1$, where we note that $UV=1$ corresponds to $r=\infty$ (whilst $r=-1$ corresponds to $UV=\pm\infty$). This may be motivated by allowing (9.5) also for $r>1$; if for such r we redefine U by $U=\exp(u)$, then (9.10) remains valid also for $r\geq1$. Furthermore, we may include $V\leq0$ in the picture, which leads to the situation described in and after the diagram below. Note that the (U,V) coordinates fail to cover all of de Sitter space; a similar construction with r replaced by $-r$ gives another coordinate system that covers the part near $r=-1$ and both systems together describe all of dS_1^2 (this will not be necessary for the Schwarzschild solution, which is easier!).



Kruskal-like diagram for (part of) de Sitter space in U-V lightlike coordinates. Region I, where $U<0$ and $V>0$, is the static patch, and region IV, where $U>0$ and $V<0$ is its mirror image (in the y-z plane). Region II, where $U>0$ and $V>0$ (below the wiggly $UV=1$ line) covers the open part of the right-hand green figure behind the x-z plane at $z<0$, whereas region III, where $U<0$ and $V<0$, covers its part at $z>0$. The part of de Sitter space left open in the right-hand green figure in front of the x-z plane is not covered by the U-V coordinates (it lies at infinity).

Returning to our Killing field vector field ∂_t for the metric, in the new coordinates we obtain

$$\partial_t = U\partial_U - V\partial_V; \tag{9.13}$$

$$g(\partial_t,\partial_t) = -8\frac{UV}{(1-UV)^2}, \tag{9.14}$$

which vanishes at $r\to 1$, as predicted. Thus the Killing field ∂_t changes from being timelike in region I (i.e. $r<1$) to being lightlike for $UV=0$ ($r=1$) to being spacelike for $U<0$, $V>0$ ($r>1$). This makes the line $r=1$, or $UV=0$, a ***Killing horizon***, a concept we will return to in §10.8; the cross $r=\pm 1$ is a ***bifurcate Killing horizon***, see Definition 10.20. From the perspective of a static observer (i.e. $r=$ constant) in the static patch $-1<r<1$, compared to Minkowski space-time the unusual situation arises that even in an infinite lifetime only signals from within the static patch will reach them, the entire rest of de Sitter space being invisible forever (in contrast, any static observer in Minkowski space-time will eventually be able to detect signals from any other physical systems anywhere in space-time). Indeed, just rotate the first picture in §5.10 by 90 degrees and you see the lightcones in de Sitter space: moving up in time, the backward lightcone does not increasingly open up, but remains confined to the static patch.

As such, the Killing horizon is also an event horizon. It can be crossed (by a non-static observer, such as a light ray or an accelerating observer), but the difference with a Schwarzschild black hole is that an observer crossing the horizon, i.e. moving from region I to region II, will not necessarily fall into a singularity, because de Sitter space has none (it is geodesically complete). Instead, the coordinate singularity $UV=1$ is simply the end of de Sitter space at infinity. As we shall see, in the Reissner–Nordström solution, cf. §9.5, the situation is again different.

9.2 The Schwarzschild solution and some of its geodesics

After this warm-up, we now state the first curved solution to the vacuum Einstein equations:[426]

$$g_S = -f(r)dt^2 + f(r)^{-1}dr^2 + r^2(d\theta^2 + \sin^2\theta d\varphi^2); \tag{9.15}$$

$$f(r) := 1 - \frac{2m}{r} = \frac{1}{r}(r - r_S). \tag{9.16}$$

This is the ***Schwarzschild metric***, defined, for the moment, for some constant $m > 0$ (see §9.5 for $m < 0$),[427] and coordinates $t \in \mathbb{R}$, $(\theta, \varphi) \in S^2$, and $r > r_S := 2m$, the ***Schwarzschild radius***. The arguably most pedagogical road towards it, which goes back to Hilbert, is as follows.[428]

1. *Staticity.* Assume: (i) $M = \mathbb{R} \times \Sigma$; (ii) coordinates adapted to this; (iii) arbitrary lapse but zero shift. Then any static solution to the vacuum Einstein equations takes the form (8.96).

2. *Spherical symmetry.*[429] Assuming $\Sigma = \mathbb{R}^3 \backslash B^3_c$ (for some $c > 0$), we may start from

$$\tilde{g} = M(r)dr^2 + r^2 d\Omega; \qquad d\Omega := d\theta^2 + \sin^2\theta d\varphi^2, \tag{9.17}$$

   where we use polar coordinates (r, θ, φ) in which the radial variable r has been normalized so as to give two-spheres S^2_r with radius r a surface areas $4\pi r^2$, as in flat space.

3. In the initial value problem,[430] staticity implies $\tilde{k} = 0$. Up to constant rescaling, the most general spatial metric $\tilde{g}$ solving the constraint (8.101) is $M(r) = f(r)^{-1}$ for some $m \in \mathbb{R}$.

4. The remaining Einstein equation (8.100) then yields $L(r) = \sqrt{f(r)}$, and hence (9.15).[431]

Note that (given the above choice of Σ), asymptotic flatness (see §8.4) follows from staticity and spherical symmetry. In §10.9 we give two other derivations of the Schwarzschild metric, namely *Birkhoff's theorem* 10.22, which derives the metric (and hence its staticity as well as its asymptotic flatness) from spherical symmetry alone, and *Israel's theorem* 10.25, which derives the metric from staticity, asymptotic flatness, and the existence of a smooth event horizon.

---

[426] A solution equivalent to this one was first found by Schwarzschild (1916). Up to differences in notation, it was stated in the above form by Droste (1916), Hilbert (1917), and Weyl (1917). Karl Schwarzschild (1873–1916) communicated his solution in a letter to Einstein dated 22 December 1915, written from the Russian front. He died on May 11, 1916, though not from the War but from the rare autoimmune skin disease *pemphigus*. Johannes Droste (1886–1963) was a PhD student of Lorentz, who did not know Schwarzschild's work but (re)discovered the solution a few months after him. Droste was a professor of mathematics at Leiden from 1930–1956. Hilbert and Weyl both cite Schwarzschild, whose solution differs from these later versions since, like Einstein at an earlier stage, he worked in unimodular coordinates (i.e. $\det(g) = -1$). See Antoci & Liebscher (2001) and Antoci (2003).

[427] In physical units, $m = GM/c^2$. The constant m equals the mass of the asymptotically flat space-time (9.15), cf. (8.108) etc. Alternatively, a static observer is described by the four-veolcity $u = f(r)^{-1/2}\partial_t$, normalized to $g(u,u) = -1$, which gives an acceleration of $\nabla_u u = mr^{-2}\partial_r$. If we replace dt^2 in the metric (9.15) by c^2dt^2, as we should in physical units, this is the same formula as in Newtonian gravity. Finally, the Schwarzschild radius $r_S = 2m$ can be found–in physical units–from Newtonian gravity as the critical radius of a gravitating ball with mass m for which the escape velocity v equals the speed of light c; indeed, one has $\frac{1}{2}c^2 = Gm/r_S$, i.e. $r_S = 2Gm/c^2$.

[428] See Hilbert (1917), of which O'Neill (1983), chapter 13, gives a modern presentation.

[429] A deeper perspective on spherical symmetry will be given in §10.9 in connection with Birkhoff's theorem.

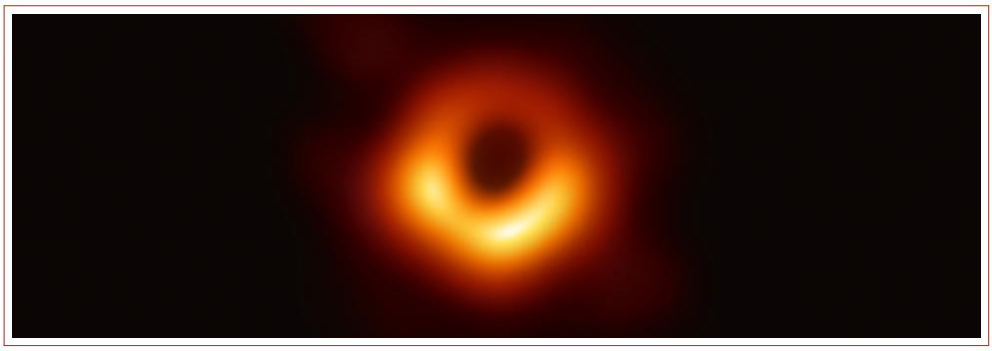
[430] As an initial-value problem the Schwarzschild case with initial data on $\Sigma = \mathbb{R}^3 \backslash B^3_c$ is unsatisfactory because as a Riemannian manifold $(\Sigma, \tilde{g})$ is incomplete, even if $c = 2m$. This will be resolved by the Kruskal solution.

[431] The simplest way to get there is to solve $\tilde{\Delta}_{\tilde{g}} L = 0$, which comes down to $(f(r)^{1/2} r^2 L'(r))' = 0$. This is solved by $L(r) = \sqrt{f(r)}$, which also solves (8.100), or by $L(r) = C$, which does not, cf. Schoen (2009), Lecture 5. Note that in the above Hilbert-style derivation the *Ansatz* "$L = L(r)$" is supposed to follow from spherical symmetry, too.

On the nose, the solution (9.15) applies to both $r > 2m$ and $0 < r < 2m$, and as we shall see in §9.3, the value $r = 2m$ is merely a coordinate singularity. Although in the present section we restrict ourselves to $r > 2m$, it is worth mentioning that $r = 0$ is a genuine singularity, both in the sense of the singularity theorems and in the sense that curvature blows up: this can be detected in a coordinate-free way through the so-called ***Kretschmann scalar***

$$R^{\rho\sigma\mu\nu}R_{\rho\sigma\mu\nu} = \frac{48m^2}{r^6}. \tag{9.18}$$

If a star has radius $R > 2m$, then its interior is modelled by some nonzero energy-momentum tensor, so that the vacuum Einstein equations to which (9.15) is a solution are only relevant for $r > R$. This is the case, for example, with our Sun. If, on the other hand, $R \leq 2m$, then the only physically stable situation to which a static solution like (9.15) could possibly apply is $R = 0$, which describes a black hole. See footnote 458. In that case, the vacuum solution (9.15) applies to both $r > 2m$ and $0 < r < 2m$. Since all black holes in the universe are believed to rotate (and hence are stationary but not static), it seems that the Schwarzschild solution for $0 < r < 2m$ does not describe anything in Nature; one would need the Kerr solution instead (see §9.6). However, one can do a few simple calculations about the metric (9.15) that are hardly changed by rotation and explain key features of the famous image of the supermassive black hole in M87:[432]



First Image of the Supermassive Black Hole in M87, revealed on April 10, 2019.[433]

A black hole obviously does not emit any radiation itself. But if it is "illuminated" (at a typical radio astronomy wavelength like 1.3 mm, so that the colors are fake), then some deflected photons may reach us and provide us with an indirect image. In the case at hand, illumination comes from a thin accretion disc whose constituents on average move around the black hole in circular geodesics and emit photons, converting gravitational energy into radiation. The aim of the following calculations is to show that the ***photon capture radius***, i.e. the radius of the central dark disc known as the ***black hole shadow***, is $a = \sqrt{27}m$, instead of $2m$ as one would find by (wrongly) identifying the disc with the interior of the black hole as defined by its event horizon. Furthermore, we will give an idea of the origin of the bright area, which is a blurred image of the ***photon sphere*** of the black hole, which is located at $r = 3m$ (from which the step to $a = \sqrt{27}m$ is straightforward geometry). The key to the structure of the black hole shadow is the existence of (unstable) circular photon orbits at radius $r = 3m$ (and no other value).[434] Perhaps paradoxically, gravitational lensing makes this radius the edge of the shadow. The instability of all circular geodesic orbits of massive particle at radiuses $r \leq 6m$ also plays a role.

[432] The Event Horizon Telescope (2019a) expects only a 4% change in a between Schwarzschild and Kerr.

[433] Source: `https://eventhorizontelescope.org`. Credit: The Event Horizon Telescope Collaboration.

[434] This was noted by Hilbert (1917)! Modern references are Luminet (1979) and Event Horizon Telescope Collaboration (2019ab). See also Misner, Thorne, & Wheeler (1973), chapter 25, and Chruściel (2020), §3.9.

To start, we describe geodesics (which as always are affinely parametrized by definition)

$$\gamma(s) = (t(s), r(s), \theta(s), \varphi(s)) \tag{9.19}$$

in the Schwarzschild metric (9.15). In the absence of off-diagonal terms, the vector fields ∂_μ are orthogonal, from which it is easy to show that the geodesic equation (3.24) becomes

$$\frac{d}{ds}(g_{\mu\mu}\dot{x}^\mu) = \tfrac{1}{2}\sum_{\nu=0}^{3}(\dot{x}^\nu)^2\partial_\mu g_{\nu\nu}, \tag{9.20}$$

where $\dot{x}^\mu = dx^\mu(s)/ds$, and $\mu = 0,1,2,3$ is fixed. Then $\mu = 0$ gives $d(f\dot{t})/ds = 0$; $\mu = 2$ gives $d(r^2\dot{\theta})/ds = r^2\sin\theta\cos\theta\dot{\varphi}^2$; and $\mu = 3$ gives $d(r^2\sin^2\theta\dot{\varphi})/ds = 0$. Hence we may set

$$f(r(s))\dot{t}(r(s)) = E; \qquad \theta(s) = \pi/2; \qquad r(s)^2\dot{\varphi}(s) = L, \tag{9.21}$$

where E and L are constants, interpreted as energy and angular momentum, respectively.[435] The case $L = 0$ gives radial motion (i.e. at constant θ and φ). If also $g_{\mu\nu}\dot{\gamma}^\mu\dot{\gamma}^\nu = 0$, then

$$t(s) = \pm(s + 2m\ln|s|) + C; \qquad r(s) = s + 2m, \tag{9.22}$$

for constant C, gives the radial lightlike geodesics, initially with $s > 0$ and hence $r > 2m$ (see §9.3 for $r < 2m$). Radial lightlike geodesics do not contribute to the black hole shadow, whereas radial timelike geodesics of massive particles do not contribute to the accretion disc, and therefore do not produce the photons in the EHT image either. Hence for understanding this image we may assume $L \neq 0$. In that case, we may invert $\varphi(s)$ to make r a function of φ instead of s.

We now use the fact that for geodesic motion the combination $g_{\mu\nu}\dot{\gamma}^\mu\dot{\gamma}^\nu$ is constant, i.e.

$$g_{\mu\nu}\dot{\gamma}^\mu\dot{\gamma}^\nu = -\lambda^2, \tag{9.23}$$

with e.g. $\lambda = 0$ for photons. Using (9.21), eq. (9.23) may be written in terms of $\mathscr{E} := E^2/L^2$ as

$$\left(\frac{1}{r^2}\frac{dr}{d\varphi}\right)^2 + V(r) = \mathscr{E}; \qquad V(r) := \left(1 - \frac{2m}{r}\right)\cdot\left(\frac{1}{r^2} + \frac{\lambda^2}{L^2}\right), \tag{9.24}$$

where in the massless case $\mathscr{E}$ is usually called $1/b^2$, with ***impact parameter*** $b = L/E$. Thus we can describe geodesic motion near a black hole as motion in a potential V, where φ plays the role of time. In fact, the second ($\mu = 1$) entry in (9.21) can also be derived from (9.24), viz.

$$\frac{1}{r^4}\frac{d^2r}{d\varphi^2} - \frac{2}{r^5}\left(\frac{dr}{d\varphi}\right)^2 = -\tfrac{1}{2}\frac{dV}{dr}. \tag{9.25}$$

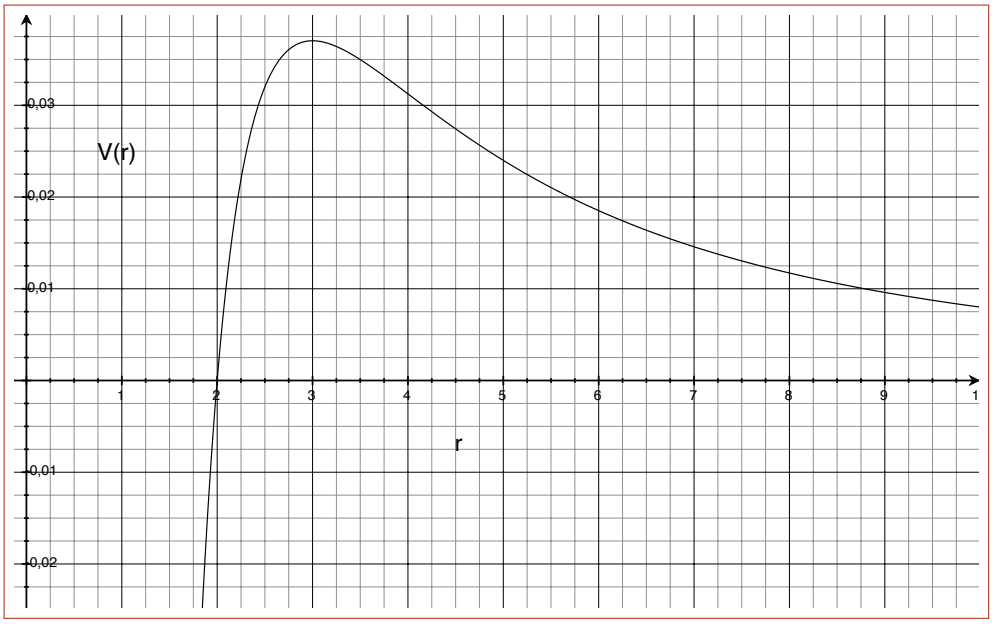
We start with the massless case $\lambda = 0$, so that the potential V in (9.24) becomes

$$V(r) = \frac{f(r)}{r^2} = \frac{1}{r^2} - \frac{2m}{r^3}. \tag{9.26}$$

This potential is plotted below. It has a maximum at $r = 3m$, at which the critical energy is

$$\mathscr{E}_c = V(3m) = 1/(27m^2). \tag{9.27}$$

[435] As we shall show systematically for the Kerr metric, see §9.6 from eq. (9.122) onwards, the conserved quantities E, L, and $\pi/2$ come from three Killing vector fields K for the Schwardschild metric, taking the form $g(\dot{\gamma}, K)$.



$V(r)$ in units of m, i.e. $V=0$ at $r=2m$ and V is maximal at $r=3m$, where $V(3m)=1/27m^2$.

- The ***photon sphere*** is the orbit $r=3m$ (i.e. $dr/d\varphi=0$) at the critical value $\mathscr{E}_c$ where V takes its maximum. Since $V'(3m)=0$, it follows from (9.25) that the circular orbit $r(\varphi)=3m$ is a geodesic. This orbit is unstable,[436] since $V''(3m)=-2/(81m^4)<0$.
- Photons with $\mathscr{E}>\mathscr{E}_c$ starting at $r>3m$ cross the barrier and fall into the black hole.
- Photons with $\mathscr{E}<\mathscr{E}_c$ are (eventually) reflected at the periastron $r_c>3m$ where $V(r_c)=\mathscr{E}$ and then increase r again (perhaps after having orbited the black hole), as in the artist impression on the next page. Such photons cannot cross the photon sphere, but depending on their energy they can come arbitrarily close to it. Almost all photons detected on earth belong to this category, which explains both the relatively sharp edge of the black hole shadow at $r=3m$, or rather $a=\sqrt{27}m$, see (9.28) below, and the bright area around it.[437]

We now explain why the apparent radius a of the black hole shadow does not equal $a=3m$ but

$$a=\sqrt{27}m. \tag{9.28}$$

Let η be the (very very very) small angle between the radial direction from us to the center of the black hole,[438] given by the vector $X=-\partial_r$, and the direction in which we see the photon.

[436] There is one exception making the photon sphere "attractive": photons whose "energy" is exactly equal to $\mathscr{E}_c$ that are not already at $r=3m$ will asymptotically spiral towards the photon sphere (and hence are invisible to us).

[437] The shadow is not absolutely black, since there is some leakage from photons coming from $2m<r<3m$. See e.g. Narayan, Johnson, & Gammie (2019), also for a general explanation of the shadow. All photons in the image of the black hole in M87 are produced as synchrotron radiation by either a hot diffuse plasma accreting onto the black hole, or a collimated plasma jet.

[438] We quote from `https://blackholecam.org/research/bhshadow/`: 'The predicted size of the shadow cast by the event horizon of the supermassive black hole at the center of our own Milky Way is about 50 microarcseconds (that is one fifty millionth of an arcsecond, which is 1/3600th of a degree!). Although super small, this angular size can actually be resolved by astronomical observations using an interferometric technique at radio wavelengths, called Very Long Baseline Interferometry or VLBI.' This makes the image by the Event Horizon Telescope an incredible technological achievement, on a par with the first detection of gravitational waves by LIGO in 2015.

This direction is the vector $Y = (dr/d\varphi)\partial_r + \partial_\varphi$. The angle is given by the usual formula

$$g(X,Y) = \sqrt{g(X,X)}\sqrt{g(Y,Y)}\cos\eta, \tag{9.29}$$

which holds in Riemannian geometry as well as it does in Euclidean geometry. Eq. (9.15) gives

$$\cos^2\eta = \frac{(dr/d\varphi)^2}{(dr/d\varphi)^2 + r^2(1-2m/r)}. \tag{9.30}$$

Eliminating $(dr/d\varphi)^2$ via (9.24) with (9.26) and $\mathscr{E} = \mathscr{E}_c$ given in (9.27) gives ***Synge's formula***

$$\sin^2\eta = \frac{27m^2(r-2m)}{r^3}. \tag{9.31}$$

Here r is our distance to the black hole, and we take $\mathscr{E} = \mathscr{E}_c$ since we want to compute the angle for the boundary of the black hole shadow, as explained above. On the other hand, Euclidean geometry as naively used by an observer at (practically) infinity in flat space-time gives

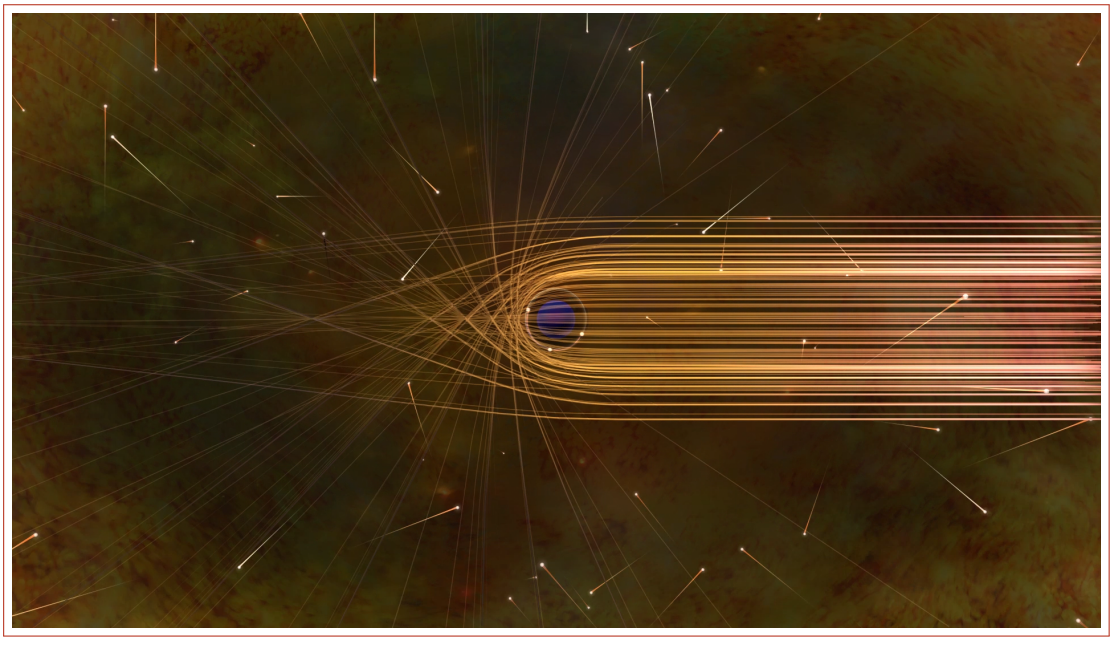
$$\frac{a}{r} = \tan\eta. \tag{9.32}$$

For very small η we have $\tan\eta \approx \sin\eta \approx \eta$. Also, in (9.31) we neglect the $2m/r^3$ term against $r/r^3 = 1/r^2$ because r is very large. Eqs. (9.31) and (9.32) then immediately yield (9.28).

Finally, we show that there are no stable circular geodesic orbits of massive particles for $r \le 6m$. First, putting $dr/d\varphi = 0$ in (9.25) gives $V'(r) = 0$, which in this case, i.e. $\lambda \neq 0$ in (9.24), unlike the massless case ($\lambda = 0$) does not lead to a unique solution but to the condition

$$r - 3m = m\lambda^2 r^2/L^2. \tag{9.33}$$

Hence $r > 3m$. Using (9.33), the stability condition $V''(r) > 0$ then becomes $r > 6m$.



Artist's impression of the paths of photons in the vicinity of a black hole. The gravitational bending and capture of light by the event horizon is the cause of the black hole shadow .[439]

[439] Source: `https://www.almaobservatory.org/en/images/photon-paths-around-a-black-hole/`. Credit: Nicolle R. Fuller/NSF.

9.3 The event horizon of Schwarzschild space-time

We now show how to cross the barrier $r = 2m$. In the coordinates used to express (9.15) this is a bit awkward,[440] since the metric is simply undefined at $r = 2m$. We resolve this coordinate singularity as in de Sitter space, see (9.6) - (9.7). We again introduce lightlike coordinates

$$u = t - r_*, \qquad t = \tfrac{1}{2}(v+u); \tag{9.34}$$

$$v = t + r_*, \qquad r_* = \tfrac{1}{2}(v-u), \tag{9.35}$$

where the new ('tortoise') radial coordinate $r_* = r_*(r)$ is defined as the solution, for $r > 2m$, of

$$\frac{dr_*(r)}{dr} = \frac{1}{f(r)}; \qquad f(r) := 1 - \frac{2m}{r}. \tag{9.36}$$

This fixes $-\infty < r_* < \infty$, corresponding to $2m < r < \infty$, up to a constant. The variables

$$x_* = (r_*/2m) - 1; \qquad x = (r/2m) - 1, \tag{9.37}$$

turn eq. (9.36) into $dx_*/dx = 1 + x^{-1}$, which for $x > 0$ is solved by $x_* = x + \ln x + C$. Hence

$$r_*(r) = r + 2m \ln \left| \frac{r}{2m} - 1 \right| = r + 2m \ln |r - 2m| - 2m \ln(2m) \tag{9.38}$$

solves (9.36). One may also solve r for r_*: for $x > 0$ we have $x = W(e^{x_*})$, where W is the ***Lambert W-function***, defined for $x > 0$ by $W(x)e^{W(x)} = x$. Up to a constant,[441] this gives

$$r(r_*) = 2m \left(W\left(e^{(r_*/2m)-1} \right) + 1 \right). \tag{9.39}$$

To interpret the coordinate r_*, note that if in a geodesic (9.19) we write the radial coordinate $r(s)$ as $r(t)$ by inverting $t(s)$, and subsequently express r in terms of r_*, from (9.21) we obtain

$$\frac{dr}{ds} = E \frac{dr_*}{dt}. \tag{9.40}$$

This relates two perspectives on radial geodesics: travellers use proper time s and undergo $s \mapsto r(s)$, whereas ***static observers***, who by definition are at rest in (r, θ, φ) and use time t,

[440] During a lecture in Paris on April 5, 1922, Hadamard asked Einstein what happened if the radius of the Sun were less than the Schwarzschild radius (Nordmann, 1922; Biezunski, 1987). After much confusion, including names (and views) of the $r = 2m$ sphere like "discontinuity" (Schwarzschild), "magic circle" (Eddington), "barrier" (Kottler), "limit circle" (Brillouin), and even "the death" (Nordmann)–we owe this information to a seminar by Dennis Lemkuhl on April 12th, 2021–at last Lemaître (1933), §11, concluded that 'The singularity of the Schwarzschild field [i.e. at $r = 2m$] is thus a fictitious singularity, analogous to that which appears at the horizon of the centre in the original form of the de Sitter universe.' Earlier, Eddington (1924) had contributed the coordinates now named after him; in fact, he did not use (u, r) or (v, r) but, with an obvious typo corrected, (t_*, r), where $t_* = t - 2m \ln |r - 2m|$, so that $u = t - r_* = t_* - r$ (up to a constant $2m \ln(2m)$). This turns the metric (9.15) into

$$ds^2 = -\left(1 - \frac{2m}{r}\right) dt_*^2 + \left(1 + \frac{2m}{r}\right) dr^2 - \frac{4m}{r} dt_* dr + r^2 (d\theta^2 + \sin^2\theta d\varphi^2),$$

which is well defined and Lorentzian for all $r > 0$. Lemaître's coordinates were different but had the same effect. Finkelstein (1958) first interpreted $r = 2m$ as an event horizon, though not using this term, which had just been introduced by Rindler (1956), who in turn did not mention the Schwarzschild solution! Penrose (1968) first rigorously put all this together. See Godart (1992), Eisenstaedt (1993), and Earman (1995), §1.2, for history.

[441] Take $x_* = \int_\varepsilon^x dy\,(1 + y^{-1})$, where $\varepsilon > 0$ solves $\varepsilon = -\ln \varepsilon$, so that in (9.38) one takes $a = 2m(1 + \varepsilon)$.

monitor $t \mapsto r_*(t)$. For example, it follows from (9.22), see also (9.66) below, that a future-directed ingoing radial lightlike geodesic approaching $r \to 2m$ from $r_0 > 2m$ takes the form

$$t(s) = s - 2m\ln(-s) + C; \qquad r(s) = -s + 2m. \tag{9.41}$$

It then follows from (9.40), in which (9.41) gives $dr/ds = -1$ and $E = 1$, that

$$r_*(t) = -t + C'. \tag{9.42}$$

Therefore, it takes $t \to \infty$ to reach $r_* \to -\infty$, which is the same as $r = 2m$ (***infinite redshift***).

For $(u,v) \in \mathbb{R}^2$, i.e. $(t, r_*) \in \mathbb{R}^2$ and hence still $2m < r < \infty$, eqs. (9.34) - (9.35) imply

$$g_S = -f(r)dudv + r^2 d\Omega, \tag{9.43}$$

where $r = r(r_*) = r(u,v)$ through (9.39) and (9.35). This shows that radial lightlike geodesics are given by constant u (outgoing) or constant v (ingoing), as in Minkowski space-time: for $r > 2m$ the former comes from the plus sign in (9.22) with $s > 0$, whereas the latter come from the minus sign with $s < 0$, as can also be directly seen from (9.38). See also (9.65) - (9.68).

To cross $r = 2m$, we use ***Eddington–Finkelstein coordinates***, in two versions: the ***ingoing*** coordinates $(v,r) \in \mathbb{R} \times (2m, \infty)$ and the ***outgoing*** ones $(u,r) \in \mathbb{R} \times (2m, \infty)$, with metrics

$$g_+ = -f(r)dv^2 + 2dvdr + r^2 d\Omega; \tag{9.44}$$

$$g_- = -f(r)du^2 - 2dudr + r^2 d\Omega. \tag{9.45}$$

These expressions suddenly make sense for any $r \in (0,\infty) \equiv \mathbb{R}^+_*$! ***Schwarzschild space-time*** is

$$M_S = \mathbb{R} \times \mathbb{R}^+_* \times S^2 \cong \mathbb{R} \times (\mathbb{R}^3 \backslash \{0\}), \tag{9.46}$$

with metric (9.44), where now $(v,r) \in \mathbb{R} \times \mathbb{R}^+_*$, and (as always) $\theta \in [0,\pi]$, $\varphi \in [0, 2\pi)$.

The reason we say that Schwarzschild space-time contains a black hole is the following.[442]

Theorem 9.1 *The* **(future) event horizon** $H^+_E = \{(v,r,\theta,\varphi) \mid r = r_S = 2m\}$ *in* M_S *is:*

1. *A smooth null hypersurface (cf. §4.6), ruled by lightlike pregeodesics (cf. Proposition 6.9).*
2. *Diffeomorphic to* $\mathbb{R} \times S^2$*;*
3. *A one-way membrane, in that fd causal cruves cannot cross* H^+_E *from* $r < r_S$ *to* $r > r_S$*.*

The last claim is predicated on a time orientation T on M_S. For $r > 2m$ this is naturally given by $T = \partial_t$, which is timelike for $r > 2m$, but (as another remarkable feature of Schwarzschild space-time) this vector becomes lightlike at $r = r_S$ and spacelike for $0 < r < r_S$. This is heuristically clear from (9.15), but since these coordinates break down at $r = r_S$ it is more precise to use (9.44), noting that $\partial_t = \partial_v$. In the absence of a geometrically natural fd timelike vector field defined throughout M_S, we therefore define time orientation via a lightlike field, namely

$$\underline{L} = -\frac{\partial}{\partial r}, \tag{9.47}$$

[442] In chapter 10 we will see that the three parts of Theorem 9.1 state general properties of (abstractly defined) event horizons: See Corollary 10.17, Proposition 10.29, and eq. (10.79), defining the event horizon, respectively.

defined in the ingoing Eddington–Finkelstein coordinates (v,r,θ,φ). This is crucial, since although the *coordinate* r is the same as in the original (t,r,θ,φ) coordinates, the *vector field* ∂_r is different.[443] As in (5.80), we then define the cone of future directed (fd) timelike vectors by

$$\mathscr{T}_x^+ = \{X_x \in \mathscr{T}_x \mid g_x(\underline{L}_x, X_x) < 0\}. \tag{9.48}$$

A timelike vector $X_x \in T_x M_S$ *is future directed (fd) iff* $g_x(\underline{L}_x, X_x) < 0$. Moving back to the original coordinates (t,r,θ,φ) it is easy to check that (9.47) - (9.48) make ∂_t timelike and fd for $r > 2m$, whereas they make $-\partial_r$ *in the original coordinates* (which is spacelike for $r > 2m$ and lightlike for $r = 2m$) timelike and fd for $r < 2m$. The disadvantage in using a lightlike field like $\underline{L}$ to define time-orientation is that one cannot define a general lightlike vector N to be fd iff $g(N,\underline{L}) < 0$ or even ≤ 0, since the former (< 0) fails for $N = \underline{L}$ whereas the latter (≤ 0) would make $N = -\underline{L}$ fd. Hence this criterion is restricted to lightlike vectors that are not proportional to $\underline{L}$.

Proof. Claim 2 follows from the coordinate definition of H_E^+, which also gives smoothness. The normal N of a hypersurface defined as a level set $f = c$ is given by $N = (df)_\sharp$, which gives $N = \partial_v + f(r)\partial_r$. Hence $g(N,N) = -f(r) + 2f(r) = f(r)$, which vanishes at $r = 2m$. Thus the normal N of H_E^+ is a lightlike vector on H_E^+ and this by definition makes H_E^+ a null hypersurface. Alternatively, since r is constant on H_E^+, the induced metric $\tilde{g}$ on H_E^+ is (9.44) at fixed $r = r_S$, i.e. $g = r_S^2 d\Omega$. This metric is degenerate, which again makes H_E^+ null, cf. §4.6.

To prove claim 3 we adopt the notation of §4.6, relabeling N as L, and have

$$L = 2(\partial_v + f(r)\partial_r); \qquad\qquad \underline{L} = -\frac{\partial}{\partial r}, \tag{9.49}$$

of which L is lightlike only on H_E^+ (where $L = 2\partial_v$), whereas $\underline{L}$ is lightlike everywhere. Note the correct normalization (6.58), which, given that (9.47) defines time orientation, implies that L is fd whenever it is causal, which is the case for $0 < r \leq r_S$. Now consider a general curve

$$c(\lambda) = (v(\lambda), r(\lambda), \theta(\lambda), \varphi(\lambda)). \tag{9.50}$$

Using (9.44), the conditions that the curve $c(\cdot)$ be timelike and future directed are, respectively,

$$g(\dot{c},\dot{c}) < 0 \qquad \Leftrightarrow \qquad 2\dot{v}\dot{r} - f(r)\dot{v}^2 + r^2(\dot{\theta}^2 + \sin^2\theta\,\dot{\varphi}^2) < 0; \tag{9.51}$$

$$g(\underline{L},\dot{c}) < 0, \qquad \Leftrightarrow \qquad \dot{v} > 0. \tag{9.52}$$

On H_E^+ we have $f(r) = 0$, which enforces $\dot{r} < 0$. This is an open condition, which by continuity also holds near H_E^+. Hence fd timelike curves must decrease r if they get near H_E^+, which means that they must either stay within the horizon $(r \leq r_S)$ or cross it from $r > r_S$.

For general causal curves: (i) eq. (9.52) should be supplemented with the additional possibility $\dot{c} = \rho\underline{L}$ for some $\rho > 0$, which clearly has $\dot{r} = -\rho < 0$; (ii) one allows zero on the right of (9.51). On the horizon, the only new case this leaves (i.e. for which $\dot{r} \not< 0$) are the so-called ***rest photons*** that have $r = r_S$ and (θ,φ) constant, and whose lightlike geodesics solve $\nabla_L L = 0$ on H_E^+ (these lightlike geodesics in fact rule the null hypersurface H_E^+). Their urge to move outward with the speed of light is exactly compensated for by the central gravitational pull, so that they are at rest at some point on S^2. Their existence does not affect the claim of the theorem. □

The proof gives us more: since $f(r) < 0$ inside H_E^+ we must have $\dot{r} < 0$ anywhere inside H_E^+ and hence any fd timelike curve within H_E^+ hits the singularity (but the rest photons do not!).[444]

[443] One may also do this in Minkowski space-time, where in (v,r,θ,φ) coordinates $(v = t + r)$, the vector field $\underline{L} = -\partial_r$ is also lightlike and fd; to see this, just note that $\underline{L} = \partial_t - \partial_r$ *in the original coordinates* (t,r,θ,φ).

[444] It would be a mistake to think that photons can somehow travel around the two-sphere $r = r_S$: as soon as $\dot{\theta}$ and/or $\dot{\varphi}$ are nonzero whilst $\dot{r} = 0$, recalling that $f(r_S) = 0$ the right-hand side of (9.51) can obviously not be zero.

9.4 The Kruskal extension of Schwarzschild space-time

The metrics (9.44) - (9.45), both defined on M_S, describe *two different space-times*, $(M_S, g_\pm)$, containing a ***black hole*** (g_+) and a ***white hole*** (g_-) respectively. The latter is a time-reversed version of the former, as follows from the fact that $(u,r,\theta,\varphi) \mapsto (v=-u,r,\theta,\varphi)$ is an isometry from (M_S,g_+) to (M_S,g_-). Note that the Schwarzschild metric is *static* for $r>2m$, whereas both (9.44) and (9.45), valid for $0<r<\infty$, are merely *stationary* and hence not time-reversal invariant. Furthermore, both space-times are extendible, and as such they will be combined into a single *inextendible* space-time. To this end, we introduce ***Kruskal coordinates*** (U,V) by:[445]

$$U = -e^{-\kappa u} = -\sqrt{\left|\frac{r}{2m}-1\right|}\, e^{\kappa(r-t)}; \qquad V = e^{\kappa v} = \sqrt{\left|\frac{r}{2m}-1\right|}\, e^{\kappa(r+t)}, \tag{9.53}$$

where $r>2m$, and κ, the so-called ***surface gravity*** at the event horizon,[446] is defined by

$$\kappa = 1/4m. \tag{9.54}$$

The pair $(u,v) \in \mathbb{R}^2$ corresponds to $t \in \mathbb{R}$ and $r>2m$, and hence to $U<0$ and $V>0$. This means that the metric (9.43) in terms of (u,v) applies; in terms of (U,V) this metric turns into

$$g_K = -\frac{32m^3}{r} e^{-r/2m} dU dV + r^2 d\Omega, \tag{9.55}$$

in which r, so far subject to $r>2m$, is regarded as a function of U and V through (9.39), (9.35), and (9.53). This dependence of r on (U,V) may (relatively) simply be stated as[447]

$$UV = \left(1-\frac{r}{2m}\right) e^{r/2m}. \tag{9.56}$$

Clearly, the metric (9.55) is well defined for $(U,V) \in \mathbb{R}^2$ as long as $r>0$. To express this constraint in terms of (U,V) we extend the transformation (9.53) as follows:

$$U = \pm\sqrt{\left|\frac{r}{2m}-1\right|}\, e^{\kappa(r-t)}; \qquad V = \pm\sqrt{\left|\frac{r}{2m}-1\right|}\, e^{\kappa(r-t)}, \tag{9.57}$$

where, using notation in which the first $\pm$ refers to U and the second to V, the signs are:

	black hole space-time	white hole space-time
$0<r\leq 2m$	+ +	- -
$r>2m$	- +	+ -

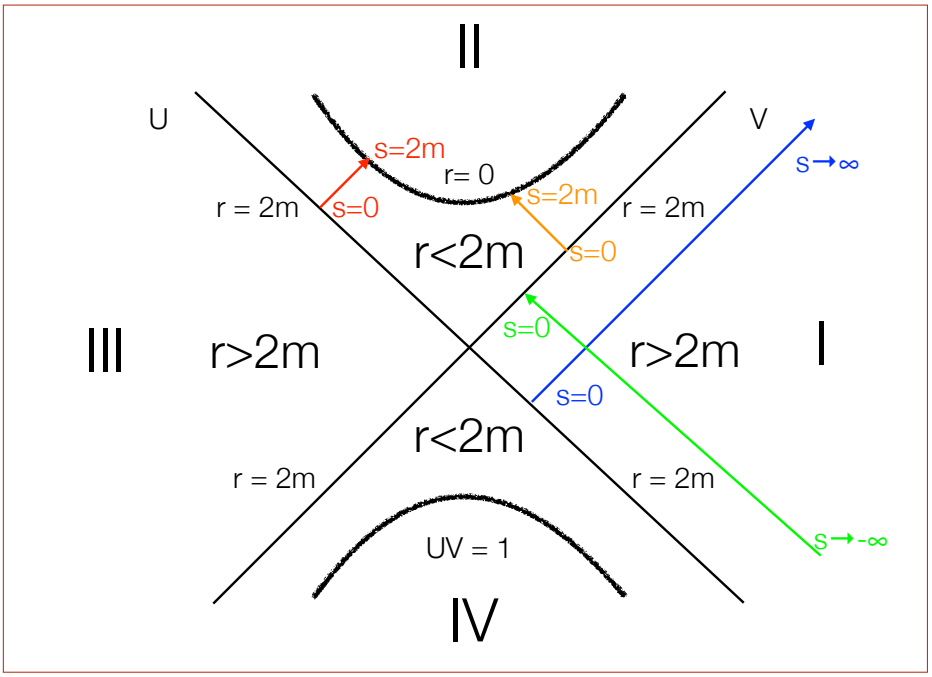
Then (9.56) remains valid for $r>0$, which gives

$$UV < 1. \tag{9.58}$$

[445] It is worth asking how these may be found. Searching for good coordinates near $r=r_S$, we approximate $f(r) \approx f(r_S)+(r-r_S)f'(r_S)+\cdots = 2\kappa(r-r_S)+\cdots$, since $f(r_S)=0$. Furthermore, near $r=r_S$ we approximate (9.38) by just keeping the logarithm, which gives $r-r_S \approx e^{2\kappa r_*}/2\kappa = e^{\kappa(v-u)}/2\kappa$. Combining these approximations gives $f(r) \approx \exp(\kappa(v-u))$, which suggests (9.53). Indeed, in terms of (U,V) the metric (9.43) may be approximated by $g \approx -\kappa^{-2} dU dV + \cdots$, which is regular near $r=r_S$. And this was the whole point of the transformation!

[446] The true significance of the surface gravity will emerge in §10.8.

[447] Following Sbierski (2018a), define $F:(0,\infty)\to(-\infty,1)$ by $F(r)=\left(1-\frac{r}{2m}\right)e^{r/2m}$, i.e. the right-hand side of (9.56). This is a homeomorphism with inverse F^{-1}, so that $r=F^{-1}(UV)$, as long as (9.58) holds.



Kruskal diagram *for the extended Schwarzschild space-time, which consists of all* $(U,V) \in \mathbb{R}^2$ *subject to* $UV < 1$*, where each point* (U,V) *is actually a two-sphere. The two wiggled lines correspond to* $r = 0 \Leftrightarrow UV = 1$ *and do not belong to the space-time; the same is true for the regions above and below these lines. The axes* $U = 0$ *or* $V = 0$ *both correspond to* $r = 2m$*. Region I is* $U < 0$ *and* $V > 0$*, etc. The Schwarzschild black hole space-time corresponds to* $U \in \mathbb{R}$ *and* $V > 0$*, and hence of regions I and II. The white hole space-time is* $U \in \mathbb{R}$ *and* $V < 0$*, and hence consists of regions III and IV. The U-V axes are rotated by* $45°$ *since radial lightlike geodesics (i.e. at constant* θ *and* φ*) correspond to straight lines* $U =$ *constant and* $V =$ *constant, as follows from* (9.55)*. The blue lightlike geodesic is* (9.69)*, the green one is* (9.70)*, the red one is* (9.71)*, and the orange one is* (9.72)*; all are future directed. All of these geodesics can clearly be extended: the blue one in the backward direction (so that it enters region IV), the green one in the forward direction (entering II), etc. This shows that the black hole space-time is extendible, as is the white hole one. But the total (Kruskal) space-time, containing both, is inextendible.*

Kruskal space-time (M_K, g_K) is given by the following space, endowed with the metric (9.55):[448]

$$M_K = \{(U,V) \in \mathbb{R}^2 \mid UV < 1\} \times S^2. \tag{9.59}$$

Schwarzschild space-time (M_S, g_+) is isometrically embedded in (M_K, g_K) as

$$M_+ = \{(U,V) \in \mathbb{R}^2 \mid U \in \mathbb{R}, V > 0, UV < 1\} \times S^2. \tag{9.60}$$

i.e. regions I and II plus the line $(U = 0, V > 0)$. For region I this follows because (9.53) transforms the metric (9.43) into (9.55), and we know that (9.43) is in turn equivalent to the Schwarzschild metric (9.44), restricted to $r > 2m$. For region II the metric (9.44), restricted to $0 < r \leq 2m$, is the unique analytic continuation of the same metric for $r > 2m$. This is also true for (9.55), and hence (9.55) must be equivalent to (9.44) in region II, too.[449]

[448] Historically, structures similar or equivalent to Kuskal coordinates and Kruskal space-time were also found by Synge, Fronsdal, and Szekeres; see Misner, Thorne, & Wheeler (1973), Box 31.1, for more information.

[449] Since regions I and III are both isometric to $\mathbb{R}^2 \times S^2$ with metric (9.43), and hence to the original $r > 2m$ Schwarzschild space-time, one might equally well realize the full space-time (M_S, g_+) as the union of II and III.

Similarly, the white hole space-time (M_S, g_-) is isometrically embedded in (M_K, g_K) as

$$M_- = \{(U,V) \in \mathbb{R}^2 \mid U \in \mathbb{R}, V < 0, UV < 1\} \times S^2, \tag{9.61}$$

and hence corresponds to regions III plus IV.[450] Let us draw the balance between the first two.

Kruskal space-time (M_K, g_K):

1. is static;
2. has a good timelike vector field defining its time orientation;
3. is globally hyperbolic;
4. is inextendible.

For the first point,[451] a simple computation shows that time translations $t \mapsto t + c$ in the original coordinates are transformed into

$$U \mapsto e^{-c/4m} U; \qquad V \mapsto e^{c/4m} V, \tag{9.62}$$

which are evidently also isometries of the metric (9.55) that preserve the condition (9.56). If t is the original time coordinate, the corresponding Killing vector field takes the simple form

$$\partial_t = \kappa(V\partial_V - U\partial_U), \tag{9.63}$$

as follows from (9.34) - (9.35) and (9.53). If we now agree that in region I the vector field ∂_t, which is timelike there, is future directed, then it follows from (9.55) that in region I, where $U < 0$, $V > 0$, $r > 2m$, both ∂_V and ∂_U are fd lightlike vector fields. Thus

$$T := \partial_U + \partial_V \tag{9.64}$$

is a globally defined fd timelike vector field that may be used to define time orientation, and which in regions I and II is compatible with the time orientation already defined by (9.47). With this time orientation, the Kruskal diagram displays what Theorem 9.1 proved, namely that the surface $r = 2m$ is an event horizon of the black hole (i.e. I + II plus their $r = 2m$ border). The event horizon at $r = 2m$ of the white hole (9.45), i.e. III + IV plus border, plays the opposite role: no fd causal curve can move from III to IV, whereas many can cross from $r < 2m$ to $r > 2m$. This follows from a similar analysis as in the proof of Theorem 9.1, with N now given by $N = +\partial_r$. For I + IV (plus border), $r = 2m$ is a one-way membrane permitting travel from IV to I but not *vice versa*, making I + IV a *white* hole. Similarly, II + III is another *black* hole.

The radial lightlike geodesics (9.22) confirm this. If we choose the (affine) parametrization such that they are all future directed,[452] we have the following four inequivalent possibilities:

$$t(s) = s + 2m\ln s + C_1; \quad r(s) = s + 2m; \quad s \in (0,\infty), t \in (-\infty,\infty), r \in (2m,\infty); \tag{9.65}$$

$$t(s) = s - 2m\ln(-s) + C_2; \quad r(s) = -s + 2m; \quad s \in (-\infty,0), t \in (-\infty,\infty), r \in (2m,\infty); \tag{9.66}$$

$$t(s) = -s + 2m\ln s + C_3; \quad r(s) = -s + 2m; \quad s \in (0,2m), t \in (-\infty,c_3), r \in (0,2m); \tag{9.67}$$

$$t(s) = s - 2m\ln s + C_4; \quad r(s) = -s + 2m; \quad s \in (0,2m), t \in (c_4,\infty), r \in (0,2m). \tag{9.68}$$

[450] It is also isometric to regions I plus IV; given (9.44) - (9.45) this would actually be the most natural identification, except that Kruskal space-time is meant to be the disjoint union of a black hole and a white hole space-time.

[451] One might think that staticity can be made explicit in ***Kruskal–Szekeres coordinates*** $t = \frac{1}{2}(V+U)$ and $x = \frac{1}{2}(V-U)$, where $(t,x) \in \mathbb{R}^2$ are constrained by $t^2 - x^2 < 1$. In terms of these, the Kruskal diagram has the usual x and t axes, and the metric is given by $g_K = \frac{32m^3}{r} e^{-r/2m}(-dt^2 + dx^2) + r^2 d\Omega$. But since r is implicitly defined by $t^2 - x^2 = (1 - r/2m)\exp(r/2m)$, cf. (9.56), this form of the metric is not manifestly t-independent either.

[452] This can be confirmed from (9.55), (9.64), and (9.69) - (9.72). For the last two, note that $1 - (s/2m) > 0$.

Here $c_3 := -z + C_3$ and $c_4 := z + C_4$ with $z := 2m(1 - \ln(2m))$. In terms of (U,V), this reads

$$U(s) = -C_1'; \qquad V(s) = C_1'' e^{s/2m} s; \qquad s \in (0,\infty); \qquad \text{(outgoing)} \qquad (9.69)$$

$$U(s) = C_2' e^{-s/2m} s; \qquad V(s) = C_2''; \qquad s \in (-\infty,0); \qquad \text{(ingoing)} \qquad (9.70)$$

$$U(s) = C_3'; \qquad V(s) = C_3'' e^{-s/2m} s; \qquad s \in (0,2m); \qquad \text{(outgoing)} \qquad (9.71)$$

$$U(s) = C_4' e^{-s/2m} s; \qquad V(s) = C_4''; \qquad s \in (0,2m), \qquad \text{(ingoing)} \qquad (9.72)$$

where all C_i' and C_i'' are positive constants (trivially computable in terms of the C_i).

For the third point, the x-axis in the Kruskal diagram above is a spacelike Cauchy surface, and the inextendibility of Kruskal space-time follows from Proposition 6.2, eq. (9.18), and a study of all geodesics in the Kruskal metric (which is not attempted here),[453] showing that all *incomplete* causal geodesics end up in the singularity at $r = 0$. Finally, the initial data problem whose MGHD is (isometric to) Kruskal space-time is asymptotically flat, albeit with two separate asymptotically flat regions of which one seems unrealistic. Hence (M_K, g_K) has good mathematical properties, but it seems not to correspond to any (known) part of our universe. In agreement with this, arguments given below suggest that Kruskal space-time cannot be the end result of an astrophysical collapse process (whereas, as we shall see, Schwarzschild can).

In contrast, Schwarzschild space-time, realized as either (M_S, g_+) or, isometrically, (M_+, g_K):

1. is merely stationary (and not static);
2. lacks a geometrically defined timelike vector field;
3. is globally hyperbolic;
4. is extendible.

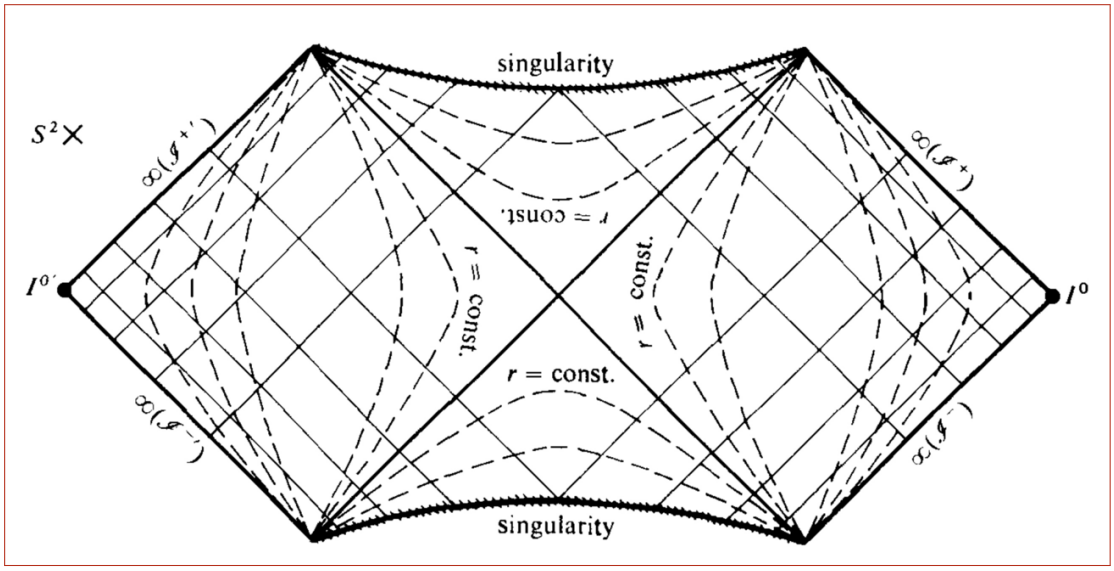
Schwarzschild space-time came about as an extension of the *static* solution (9.15), defined for $r > 2m$, to all values $0 < r < \infty$, but this extension is no longer static because of the off-diagonal terms in the metric (9.44). However, its maximal extension, i.e. Kruskal space-time, is once again static: adding a white hole to a black hole restores symmetry under time reversal.[454] As to the second point, in compensation (M_S, g_+) does have a natural *lightlike* vector field, viz. (9.47); see the proof of Theorem 9.1. For the third, Schwarzschild is globally hyperbolic, but any underlying Cauchy surface Σ would have to extend into both regions I and II in the Kruskal diagram drawn above (it cannot be restricted to region I since e.g. the red lightlike geodesic just described and drawn would not cross it). In that case Σ would still carry ***complete initial data***; that is, the Riemannian three-manifold $(\Sigma, \tilde{g})$ is geodesically complete. Region I is also globally hyperbolic by itself, with for example the positive x-axis as a Cauchy surface Σ_I. But here the initial data are incomplete because many geodesics end at $r = 2m$, and the resulting space-time is once again extendible. In this case, the $r = 2m$ hypersurface H_E^+ acts also as a future Cauchy horizon $H_C^+ = \partial D^+(\Sigma_I) \backslash \Sigma_I$ for Σ_I, seen as a wannabe Cauchy surface for the extension (M_S, g_+), cf. (5.182) and (10.86), which then coincides with the future event horizon.

[453] See for example O'Neill (1983), chapter 13 or Plebański & Krasiński (2006), chapter 14. The crucial result is Proposition 13.36 in O'Neill (1983), which states that an inextendible timelike geodesic $\gamma : I \to M_K$ in (M_K, g_K) is incomplete iff $r\gamma(s) \to 0$ as the affine parameter s approaches a finite endpoint of I, with Corollary 13.37 to the effect that Kruskal space-time is (causally) incomplete and inextendible.

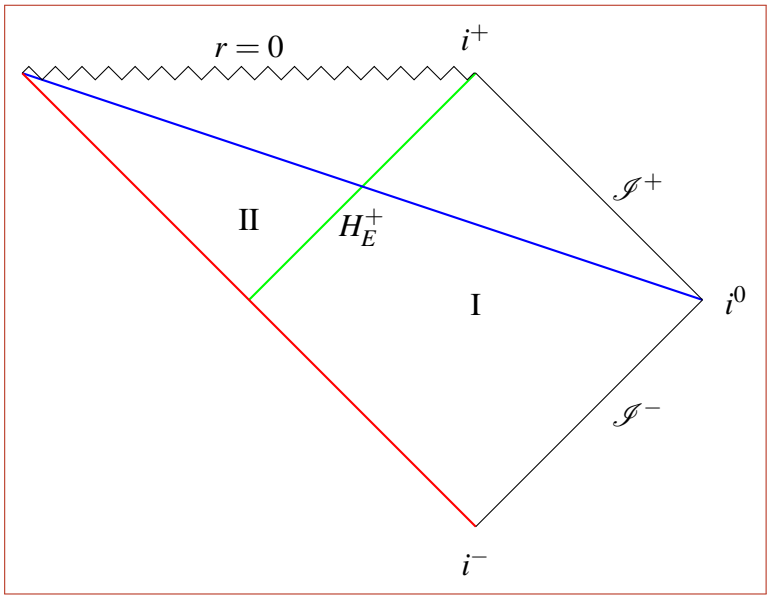
[454] Recall that a metric is static iff it is stationary for a hypersurface-orthogonal Killing vector field, which is the case iff it is stationary and time-reversal invariant (in the flow parameter of the said Killing field), see §8.4.

This reflects the extendibility of Schwarzschild space-time: for example, the radial lightlike geodesic (9.69) can be extended to negative values of s and then describes a photon moving from IV to I in finite affine parameter "time".[455] Similarly, the radial lightlike geodesic (9.71) can be extended to $s < 0$ and then describes a photon moving from IV to II.

For those who are familiar with this technique (or jump to §10.2),[456] *we now give the Penrose diagrams of both Kruskal and Schwarzschild space-time, the former in Penrose's own hand:*



One of the first Penrose diagrams: 'The Kruskal picture with conformal infinity represented.'[457]



Penrose diagram for Schwarzschild space-time $(M_+, g_K) \cong (M_S, g_+)$. The green line represents the event horizon H_E^+ at $r = 2m$. The blue line represents a Cauchy surface. The red line marks the end of the diagram; it does not (even) lie in the conformal completion $(\hat{M}_+, \hat{g}_K)$.

[455] The reason we will not notice this even if white holes exist is that according to the description (9.65), it would require extending our time t beyond minus infinity, i.e. the "beginning of time", to see it.

[456] For the moment, just note that (i) through a conformal transformation, infinity has been brought forward so as to become a boundary at some finite distance; (ii) the causal structure is the same as in Minkowski space-time.

[457] Taken from Penrose (1968), p. 208, Fig. 37. See the Introduction for comments on his style. See also §10.3.

On the other hand, we shall see at the end of this section that Schwarzschild space-time *can* result from a realistic collapse process, and although all known black holes in the universe seem to be rotating and described by the Kerr metric (with poorly known angular velocities, that is), the Schwarzschild metric may be sufficiently close to these to call it physically realistic.

In conclusion, Kruskal space-time has good mathematical features, but is physically awkward, whereas Schwarzschild space-time has exactly the opposite features. Perhaps it is a mistake to regard the latter as the "hydrogen atom of GR", as many textbooks suggest.

At the origin $U = V = 0$ of the Kruskal diagram (where $r = 2m$ and t is undefined) the event horizon of the black hole coincides with the one of the white hole. This point (which is really a two-sphere whose abstract structure is that of a *bifurcation surface*, see §10.8) is called an ***Einstein–Rosen bridge***, which later came to be seen as a special case of a ***wormhole***. This bridge connects region I to region IV, but one cannot cross it since this would require spacelike (i.e. superluminal) travel; even any (fd) timelike or lightlike deviation from it would cause the traveler to fall into the black hole singularity. Nonetheless, one can study its geometry at some fixed value of t, i.e. as part of a slice of constant U/V), which turns out to be quite interesting. We restrict ourselves to the original description of the bridge by Einstein & Rosen (1935) themselves, since apart from some use in science fiction the idea seems to be of historical value only.

In terms of the coordinate

$$u = \sqrt{r - 2m}, \tag{9.73}$$

the $r > 2m$ part of the Schwarzschild metric is

$$g = -\frac{u^2}{u^2 + 2m} dt^2 + 4(u^2 + 2m)du^2 + (u^2 + 2m)^2 d\Omega. \tag{9.74}$$

Although $u \geq 0$ initially, this makes sense for any $u \in \mathbb{R}$ and as such the solution describes the exterior regions I and IV in the Kruskal diagram. The area of any two-sphere at fixed u is

$$A(u) = 4\pi(2m + u^2)^2. \tag{9.75}$$

This function obviously takes a minimum at $u = 0$, i.e., $r = 2m$, and increases for larger $|u|$. At fixed θ, where the spheres are circles, one may then draw the bridge as a two-sided trumpet.

We return to the physical origin of Schwarzschild space-time, in the sense that it may be the final state of a stellar collapse. To this, end, the oldest and simplest generally relativistic model is due to Oppenheimer and Snyder (1939), whose paper played an important role towards the acceptance of black holes, at a time where the mathematical possibility was clear but Einstein, Eddington, and many other opinion leaders believed that they were idealizations and that some physical mechanism would block their actual occurrence in nature.[458] This model describes the collapse of a spherically symmetric permeating dust cloud, whose energy-momentum tensor within the cloud is given by (7.70), whilst $T_{\mu\nu} = 0$ outside the cloud, which is taken to be a ball

[458] Very briefly: light stars retire as white dwarfs, in which nuclear burning has ended and inward gravitational pressure is stopped by a degenerate electron gas. In 1931, Chandrasekhar discovered that this only works for masses $M \leq 1.46 M_\odot$, where $M_\odot$ is the solar mass. Heavier stars collapse into neutron stars (typically after a supernova explosion), but also these have an upper mass, as first suggested by Oppenheimer & Volkoff (1939); the current bound is about $2.3 M_\odot$. Heavier stars have nothing to stop gravitational collapse and unless they get rid of some of their mass/energy they must collapse into a black hole. See e.g. Misner, Thorne, & Wheeler (1973), Joshi (2007), Lasky (2010), or Weinberg (2020) for the relevant astrophysics, and the references in footnote 270, as well as Longair (2006), for details on the history. Our brief mathematical treatment below is based on Alford (2020).

in $\mathbb{R}^3$ with radius R. This model has only one free parameter, namely the total mass m (initially of the collapsing matter, eventually of the black hole). At any point in time t one has

$$m = (4\pi/3)R^3\rho, \tag{9.76}$$

where the choice of R ($> 2m$) reflects the choice of the origin $\tau = 0$ of (proper) time. Given this choice, define $\tau_0 = \frac{1}{3}\sqrt{2R^3/m}$, in terms of which ***Oppenheimer–Snyder space-time*** is given by

$$M = \mathbb{R}^4 \setminus \{(\tau, r, \theta, \varphi) \mid \tau \geq \tau_0, r = 0\}; \tag{9.77}$$

$$g_{OS} = -\left(1 - \frac{2m}{r}\right) d\tau^2 + 2\sqrt{\frac{2m}{r}} d\tau\, dr + dr^2 + r^2 d\Omega; \qquad (r \geq r_b(\tau)); \tag{9.78}$$

$$g_{OS} = -\left(1 - \frac{2mr^2}{r_b^3}\right) d\tau^2 + 2\sqrt{\frac{2mr^2}{r_b^3}} d\tau\, dr + dr^2 + r^2 d\Omega; \; (r < r_b(\tau)), \tag{9.79}$$

where, compared to the original coordinates (t, r, θ, φ), we have $\tau = t + g(r)$, where $g(r)$ solves

$$\frac{dg(r)}{dr} = \frac{\sqrt{2mr}}{r - 2m}. \tag{9.80}$$

Indeed,[459] under this coordinate transformation (9.78) is equivalent to (9.15). Furthermore, in (9.79) the time-dependent radius of the star $r_b = r_b(\tau)$ is defined in terms of $R = r_b(\tau = 0)$ by

$$r_b(\tau) = \left(R^{3/2} - \frac{3\tau}{2}\sqrt{2m}\right)^{2/3}. \tag{9.81}$$

Hence $r_b(\tau_0) = 0$, which means that, as suggested by (9.77), the collapse ends at $\tau = \tau_0$ and hence for all $\tau \geq \tau_0$ one has the Schwarzschild solution. Another critical time is $\tau = \tau_1$, at which $r_b(\tau_1) = 2m$ and hence the star implodes through its Schwarzschild radius. Using reduced radii

$$\tilde{r} = r/2m; \qquad \tilde{r}_b = r_b/2m; \qquad \tilde{R} = R/2m, \tag{9.82}$$

the relevant quantities are given by

$$\tilde{r}_b(\tau) = \left(\tilde{R}^{3/2} - \frac{3\tau}{4m}\right)^{2/3}; \tag{9.83}$$

$$\tau_0 = \frac{4}{3} m \tilde{R}^{3/2}; \qquad \tau_1 = \frac{4}{3} m (\tilde{R}^{3/2} - 1); \qquad \tau_0' = \frac{4}{3} m \left(\tilde{R}^{3/2} - \frac{27}{8}\right), \tag{9.84}$$

where τ_0' is the earliest time at which the quantity $\tilde{r}(\tau)$ defined below vanishes. The event horizon H_E^+ can be computed from the fact that, at any fixed angle (θ, φ), the fd "outgoing" (but bouncing) radial lightlike geodesic that passes through $(r = 2m, \theta, \varphi)$ at $\tau = \tau_1$ is given by

$$\tilde{r}(\tau) = \tilde{r}_b(\tau)(3 - 2\sqrt{\tilde{r}_b(\tau)}). \tag{9.85}$$

This takes its maximum $\tilde{r} = 1$ at $\tau = \tau_1$ and, constrained by $\tilde{r} \geq 0$, has two zeros at τ_0 and τ_0'; clearly, $\tau_0' < \tau_1 < \tau_0$. Together with (9.77) this gives the location of the event horizon as

$$H_E^+ = \{(\tau, r, \theta, \varphi) \in M \mid (\tau_0' \leq \tau \leq \tau_1, \tilde{r} = \tilde{r}(\tau)) \vee (\tau \geq \tau_1, r = 2m)\}. \tag{9.86}$$

[459] The metric is only piecewise smooth but satisfies appropriate junction conditions at $r = r_b(\tau)$.

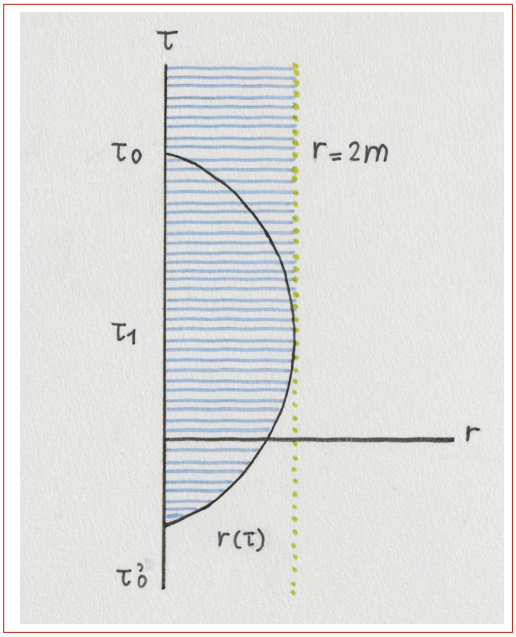
Indeed, this definition is such that some point $x = (\tau, r, \theta, \varphi)$ lies inside the horizon, in that:

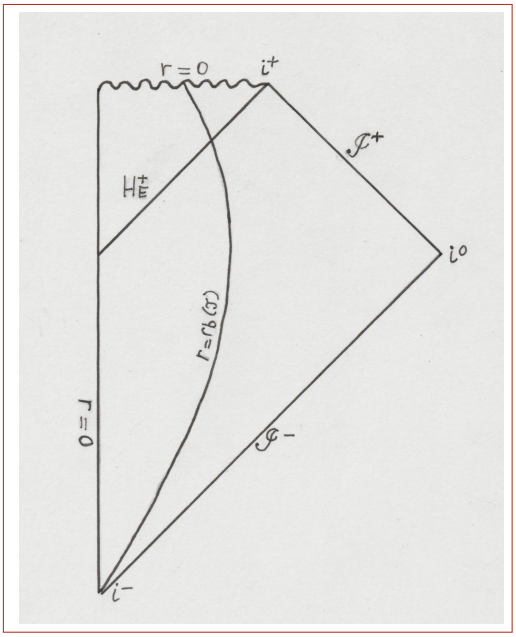
$$\tau_0' \leq \tau \leq \tau_1 \text{ and } \tilde{r} \leq \tilde{r}(\tau) \text{ or } \tau \geq \tau_1 \text{ and } \tilde{r} \leq 1, \ (i.e.\ r \leq 2m),$$

iff there are no points

$$y = (\tau', r', \theta', \varphi') \in J^+(x) \tag{9.87}$$

for which $\tau' \geq \tau_1$ and $r' > 2m$. This, in turn, means that future null infinity $\mathscr{I}^+$ cannot be reached from x via a lightlike curve (or any causal curve); we will formalize this later.[460] Specifically, lightlike geodesics starting anywhere at any time $\tau < \tau_0'$ reach infinity, whereas those starting at some $\tau \geq \tau_0'$ must start at some $\tilde{r} > \tilde{r}(\tau)$. The geodesics (9.85) demarcate between these two classes. The situation is illustrated in the pictures, which say more than the formulae:





Left picture: τ-r diagram of the Oppenheimer–Snyder space-time. The green area is the interior of the black hole and its boundary. The event horizon H_E^+ is initially the blue geodesic (9.85), and from $\tau = \tau_1$ onwards it is the line $r = 2m$. Any inextendible fd lightlike curve leaving outside the green area will eventually reach future null infinity $\mathscr{I}^+$. Any fd causal curve leaving within the grey area will stay there and any such fd causal geodesic necessarily fall into the singularity. Picture drawn by Edith de Jong.

Right picture: Penrose diagram of the Oppenheimer–Snyder space-time, very slightly adapted from Alford (2020), redrawn by Edith de Jong. The curved line shows the evolution of the radius of the star; the 45° line marked H_E^+ is the event horizon. This diagram combines features of the corresponding diagrams for Minkowski space-time (cf. §10.2) and for Kruskal space-time (given earlier in this section).

Although the romanticism has been taken out of it, one cannot deny the physical and mathematical improvement over Schwarzschild (or Kruskal) space-time: Oppenheimer–Snyder space-time is geodesically incomplete only at $r = 0$, where it has the same curvature singularity as Schwarzschild (the vertical $r = 0$ line belongs to the space-time until $\tau = \tau_0$), and hence it is inextendible–so no need for white holes. Finally, it has a complete, asymptotically flat initial value problem: any hypersurface $\tau =$ constant at $\tau < \tau_0'$ is a space-like Cauchy surface.

[460] See §10.3. The black hole area will formally be defined as $M \backslash J^-(\mathscr{I}^+)$, so that the event horizon is $H_E^+ = \partial(M \backslash J^-(\mathscr{I}^+))$. This also gives the even horizon H_E^+ of the Schwarzschild solution, as well as the horizons H_E^+ in the Reissner–Nordström and Kerr solutions to come in the next two section.

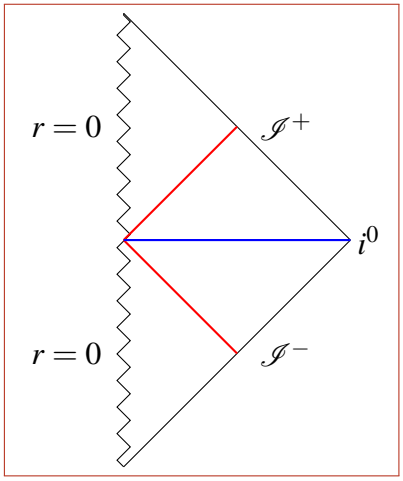
9.5 The Reissner–Nordström solution

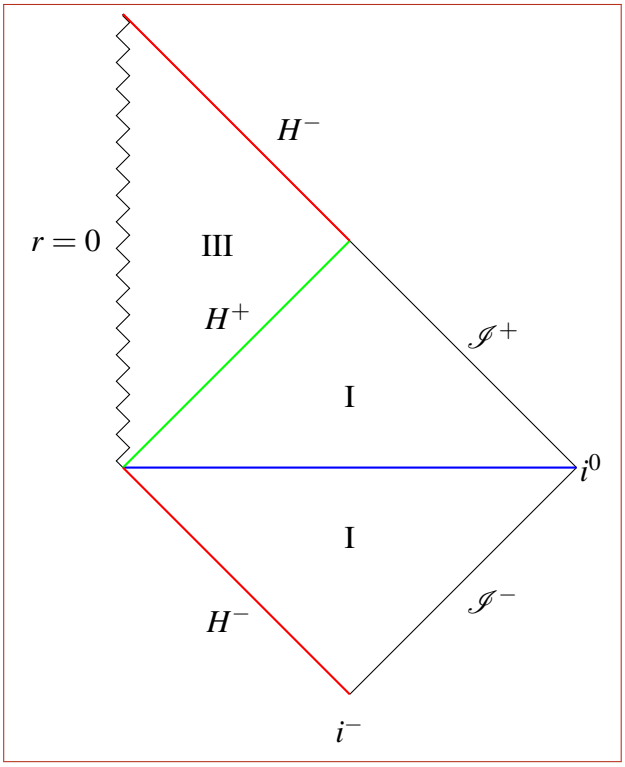
Reissner (1916) and Nordström (1918) independently extended the Schwarzschild solution to the electrovac case where the central body is electrically charged and one continues to assume spherical symmetry.[461] This requires a nonzero energy-momentum tensor (7.84) in which $F_{\mu\nu}$ comes from the potential $(A_0 = -e/r, A_i = 0)$. A lengthy calculation gives the metric

$$g_{RN} = -h(r)dt^2 + h(r)^{-1}dr^2 + r^2 d\Omega \tag{9.88}$$

$$h(r) := 1 - \frac{2m}{r} + \frac{e^2}{r^2} = \frac{1}{r^2}(r - r_+)(r - r_-); \qquad r_\pm = m \pm \sqrt{m^2 - e^2}, \tag{9.89}$$

where we assume $m > 0$ and $|e| \le m$ only in rewriting $h(r)$ as $(r - r_+)(r - r_-)/r^2$. For $e > m > 0$ we have $h(r) > 0$ and the metric (9.88), defined for all $t \in \mathbb{R}$ and $r > 0$, turns out to be inextendible. Other parameter values require new coordinates near both $r = r_\pm$, see below.

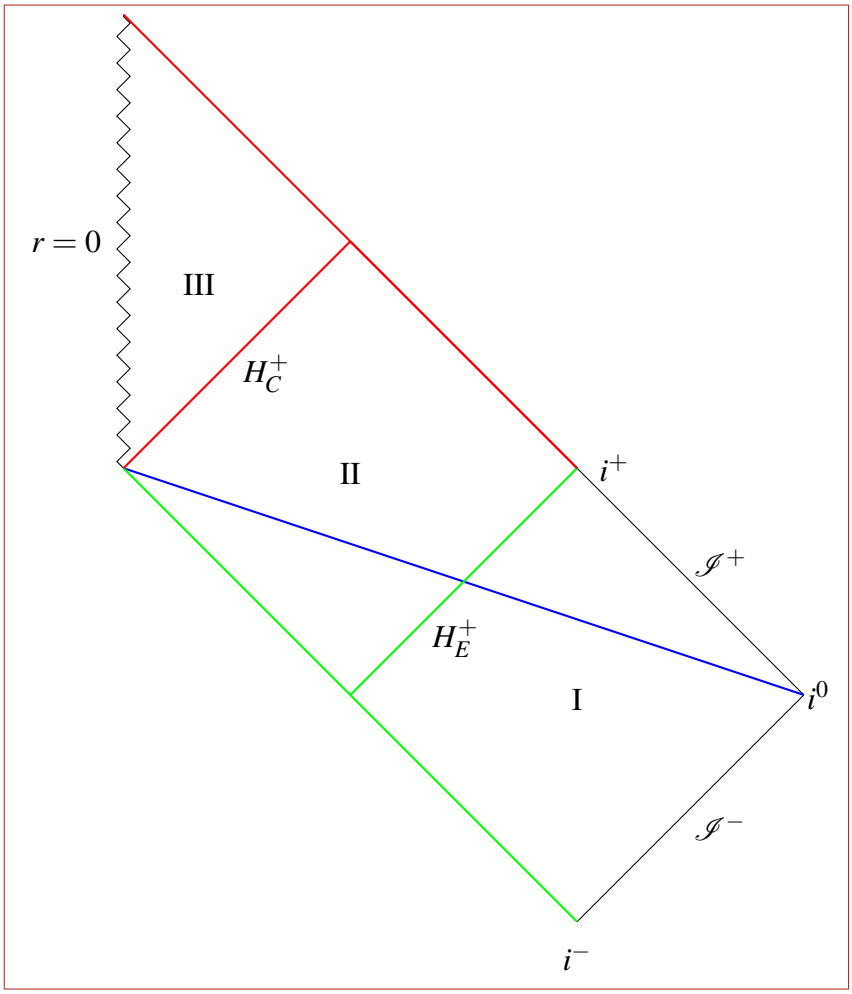




Left: *Penrose diagram for the Reissner–Nordström solution with $|e| > m > 0$ or the Schwarzschild solution with $m < 0$. The analogy with the corresponding diagram for Minkowski space-time (see §10.2) is highly misleading, since in the solutions just mentioned $r = 0$ is a naked (timelike) singularity, whereas in the Minkowski case it is a coordinate singularity. The Reissner–Nordström metric has a (naked and timelike) singularity at $r = 0$ and lacks an event horizon. It does have a Cauchy horizon, drawn in red, for the wannabe Cauchy surface drawn in blue. See §10.6 for details.*

Right: *Penrose diagram for the (unextended) Reissner–Nordström solution with $|e| = m > 0$. The singularity at $r = 0$ is shielded by a future event horizon at $r = m$, drawn in red, which at the same time is a future Cauchy horizon for the wannabe Cauchy surface drawn in blue, whence we write $H_C^+ = H_R^+ = H_C^+$. The singularity is* timelike *(the $m > 0$ Schwarzschild singularity is* spacelike*). The two red lines marked H^- are boundaries, but in the extensions discussed below they will be past event and Cauchy horizons.*

[461] There is a Birkhoff-style derivation of this metric that only requires spherical symmetry (Hoffmann, 1932ab).



Penrose diagram for the (unextended) Reissner–Nordström solution with $0 < |e| < m$*, as redefined in ingoing Eddington–Finkelstein coordinates* (v,r)*. This time there is both an event horizon* H_E^+ *at* $r = r_+$*, viz. the green center–NE line, and a Cauchy horizon* H_C^+ *at* $r = r_- < r_+$ *for say the wannabe Cauchy surface in blue, drawn as the red center–NE line. The other green and red lines are event and Cauchy horizons, respectively, for extensions of the space-time.*[462]

Key intuition about this metric comes from the Penrose diagrams above. Although charged black holes probably do not exist, pedagogically the Reissner–Nordström solution is a useful intermediate case between Schwarzschild and Kerr.[463] There are three very different regimes, which however have the same curvature singularity at $r = 0$, which is given by,[464] cf. (9.18),

$$R^{\rho\sigma\mu\nu}R_{\rho\sigma\mu\nu} = \frac{48m^2}{r^{12}}(r^6 - 2me^2r^5 + \tfrac{7}{6}e^4r^4). \tag{9.90}$$

[462] See (9.94) below for (v,r), The coordinate transformations leading to the conformal completion implicit in this Penrose diagram are given in Hawking & Ellis (1973), p. 157, but one may also use Penrose's formulae (10.72) and (10.73) for (U_+,V_+) instead of (U,V), as defined in (9.103) below. In any case, the green SE–NW line corresponds to $v = -\infty$, whereas the green SE–NW line upt to H_C^+ corresponds to $v = \infty$; at H_C^+ the (v,r) coordinates break down, as explained in the main text. One has a similar Penrose diagram for the outgoing Eddington–Finkelstein coordinates (u,r), which contains a white hole, see e.g. Poisson (2004), §5.2.3. These can be combined, see below.

[463] An exhaustive study of the Reissner–Nordström space-time and its properties may be found in Chandrasekhar (1983), chapter 5. For briefer treatments see also Graves & Brill (1960), Carter (1973), Simpson & Penrose (1973), Hawking & Ellis (1973), §5.5, Poisson (2004), §5.2, and Plebański & Krasiński (2006), §5.2.3.

[464] See Henry (2000), who even computes the Kretschmann scalar for the Kerr–Newman metric.

- $m < 0$ or $0 < m < |e|$. Then $h(r) > 0$ and the metric (9.88) is non-singular except at $r = 0$. This case is similar to Schwarzschild with $m < 0$. The metric (9.88) is defined for all $0 < r < \infty$ and the space-time is inextendible. With $T = \partial_t$ as the obvious time orientation (which for $|e| > m$ stays timelike for all $r > 0$, as opposed to $m > 0$ Schwarzschild), there are both future-directed past-incomplete timelike curves emanating from the singularity and future-directed future-incomplete timelike curves crashing into it, so that the singularity behaves like a point omitted from some (in fact many) double cone(s) $J(x,y)$. Thus a space-time with a timelike singularity cannot be globally hyperbolic.[465]
- $0 < m = |e|$, called ***extremal***, see below. We will treat this as a limiting case of:
- $0 < |e| < m$. Though all cases are unphysical, this one is "relatively realistic".

In the last two cases we have to deal with zeros of $h(r)$, where the metric (9.88) breaks down. As in the Schwarzschild case with $m > 0$, this is resolved by turning to better coordinates, and indeed, we proceed in almost the same way. The tortoise coordinate r_* now solves $dr_*/dr = h(r)^{-1}$, so that, with the simplest integration constant, eq. (9.38) is replaced by

$$r_* = r + \frac{r_+^2}{(r_+ - r_-)} \ln(r - r_+) - \frac{r_-^2}{(r_+ - r_-)} \ln(r - r_-); \qquad (0 < |e| < m). \tag{9.91}$$

Up to a constant $2m\ln(2m)$, this reduces to (9.38) if $r_- = 0$; we still have the boundary condition $\lim_{r \downarrow r_+} r_*(r) = -\infty$. The Schwarzschild surface gravity $\kappa = 1/4m$ is now replaced by

$$\kappa_+ = \tfrac{1}{2} h'(r_+) = \frac{r_+ - r_-}{2r_+^2} = \frac{\sqrt{m^2 - e^2}}{r_+^2}. \tag{9.92}$$

Thus the metric with $|e| = m > 0$ has zero surface gravity, making it an ***extremal black hole***.

The counterparts of the Schwarzschild(ish) metrics (9.43) and (9.44) - (9.45) are given by

$$g_{RN} = -h(r)dudv + r^2 d\Omega; \tag{9.93}$$

$$g_{RN} = -h(r)dv^2 + 2dvdr + r^2 d\Omega; \tag{9.94}$$

$$g_{RN} = -h(r)du^2 - 2dudr + r^2 d\Omega, \tag{9.95}$$

where u and v are defined as in (9.34) - (9.35), and as before $r = r(u,v)$ via (9.35) and the inverse of the counterpart of (9.38). Taking (9.94) to define the metric g_{RN}, we may now define ***Reissner–Nordström space-time*** as (M_{RN}, g_{RN}) where the manifold M_{RN} is the same as the Schwarzschild manifold M_S defined in (9.46), and time orientation is given by declaring that the lightlike vector field (9.47) be future directed, just as in the Schwarzschild case. Under the map

$$(v, r, \theta, \varphi) \mapsto (u = -v, r, \theta, \varphi), \tag{9.96}$$

this "ingoing" space-time is isometric to the "outgoing" one based on the same manifold, but using the metric (9.95), and $+\partial_r$ for time orientation.

[465] In Penrose's (1979) terminology, this makes the singularity ***timelike***. In addition, it is ***naked*** in being visible far away, since it is not covered by an event horizon. In contrast, the Schwarzschild singularity for $m > 0$ is ***spacelike*** and is covered by an event horizon. A singularity is spacelike/timelike/lightlike iff it has these properties in a Penrose diagram. In this case, where the singularity is located at $r = 0$, spacelike also means that for small enough $\varepsilon > 0$ the hypersurface $r = \varepsilon$ is spacelike. This is the case for Schwarzschild with $m > 0$, since its normal ∂_r is timelike for $0 < r < 2m$, whereas this normal is spacelike for all cases of Reissner–Nordström, making the singularity timelike, see Definition 4.15. These things come to a head in cosmic censorship, see §10.4.

The main properties of Schwarzschild space-time are arguably that (i) it has a spacelike curvature singularity at $r = 0$ (which makes it geodesically incomplete), which (ii) is covered by an event horizon, as expressed in Theorem 9.1. Reissner–Nordström has this singularity, too, but if $0 < |e| < m$ it is covered by *two* event horizons (and this also turns out to make it timelike):

Theorem 9.2 *If $0 < |e| < m$, the sets $H_\pm = \{(v,r,\theta,\varphi) \mid r = r_\pm\} \subset M_{RN}$ (at which $h = 0$) are null hypersurfaces diffeomorphic to $\mathbb{R} \times S^2$. Each $H_\pm$ acts as a one-way membrane towards smaller values of r. If $|e| = m$, then $H_+ = H_- = H$, which has the same properties as each $H_\pm$.*

Proof. The proof of Theorem 9.1 is easily adjusted, cf. the proof of Theorem 9.3 for details. □

This makes Reissner–Nordström space-time totally different from the Schwarzschild one, *even in the extremal case with a single event horizon.* The key properties for $0 < |e| < m$ are:

1. The *outer* horizon $H_+ = H_E^+$ is *the* event horizon, since it is the boundary inside which future (null) infinity can no longer be reached; it is the analogue of H_E^+ in Theorem 9.1.
2. The *inner* horizon $H_- = H_C^+$ is a *Cauchy horizon* for ***wannabe Cauchy surfaces*** (cf. Definition 5.36). Cauchy surfaces do not exist and (M_{RN}, g_{RN}) is not globally hyperbolic.[466]
3. The singularity at $r = 0$ is timelike and repulsive (except for radial lightlike geodesics).
4. The maximally extended space-time has an infinite number of regions (and singularities).

In the extremal case $0 \neq |e| = m$ all this remains true: although there is a single event horizon in that case, it plays the role of both the event horizon H_E^+ of Schwarzschild space-time *and* of a Cauchy horizon (which is absent in Schwarzschild space-time). For $|e| > m > 0$, finally, only property 3 remains, but as a relic of no. 2 also that case is not globally hyperbolic.[467]

Except for the first, which is Theorem 9.2, we will not prove these points (which could be done by studying all geodesics and causal curves), but just argue for them, and relate them.

The most intuitive point is 3. In the coordinates (t,r,θ,φ) the vector field $R = -\partial_r$ is spacelike for $r > r_+$, lightlike at $r = r_+$, and timelike at $r_- < r < r_+$. In this region R is future directed and hence r must decrease, so that $r = r_-$ is reached, and crossed. If $0 < r < r_-$, then R becomes spacelike once again, whereas the Schwarzschild-R only changes its causal nature once, namely when crossing the single event horizon H_E^+, and so R is timelike as $r \to 0$. This makes the Schwarzschild singularity spacelike (as the normal vector to the $r = \varepsilon$ hypersurface for small $\varepsilon > 0$ is timelike) and unavoidable (since fd timelike curves *must* decrease r), whereas the Reissner–Nordström singularity is timelike, since exactly the opposite causal situation reigns.

This suggests that the Reissner–Nordström singularity at $r = 0$ can be avoided; what's more, a fd timelike geodesic cannot even reach it because it is repelled! We only show this for radial geodesics γ (which by their very nature should have the best chance of hitting the singularity), but it is true in general. Taking (θ,φ) constant, we parametrize $\gamma(s) = (v(s), r(s))$ affinely such that $g(\dot\gamma,\dot\gamma) = -1$, where $\dot\gamma = d\gamma/ds$ as usual. In the ingoing (v,r,θ,φ) coordinates this gives

$$h\dot v^2 - 2\dot v\dot r = 1. \tag{9.97}$$

Furthermore, since $\partial_t = \partial_v$ is a Killing vector, the energy $E = -g(\dot\gamma, \partial_t)$ takes the constant value

$$E = h\dot v - \dot r. \tag{9.98}$$

[466] In §10.7 we will see that Cauchy horizons are always ruled by lightlike geodesics, sharpening Theorem 9.2.

[467] As one can infer from its Penrose diagram: no wannabe Cauchy surface, drawn as a more or less horizontal line, is hit by a future inextendible timelike curve lying above it that hits the singularity.

Combining (9.97) - (9.98) we see that, similarly to (9.24), the motion is controlled by

$$\dot{r}^2 + h(r) = E^2. \tag{9.99}$$

That is, $h(r)$ acts like a potential. Since $h(r) \approx e^2/r^2$ for $r \to 0$, this gives a strong repulsion. On the other hand, incoming fd radial lightlike geodesics are simply given by constant (θ, φ) and

$$v = C_1; \qquad\qquad r(s) = -s + C_2, \tag{9.100}$$

where C_1 and $C_2 > 0$ are constants. Since we now have $g(\dot{\gamma}, \dot{\gamma}) = 0$, eq. (9.97) is $0 = 0$ whilst (9.98) is $E = 1$, from which nothing can be concluded. Yet (9.100) gives $r(s) \to 0$ as $s \to C_2$.

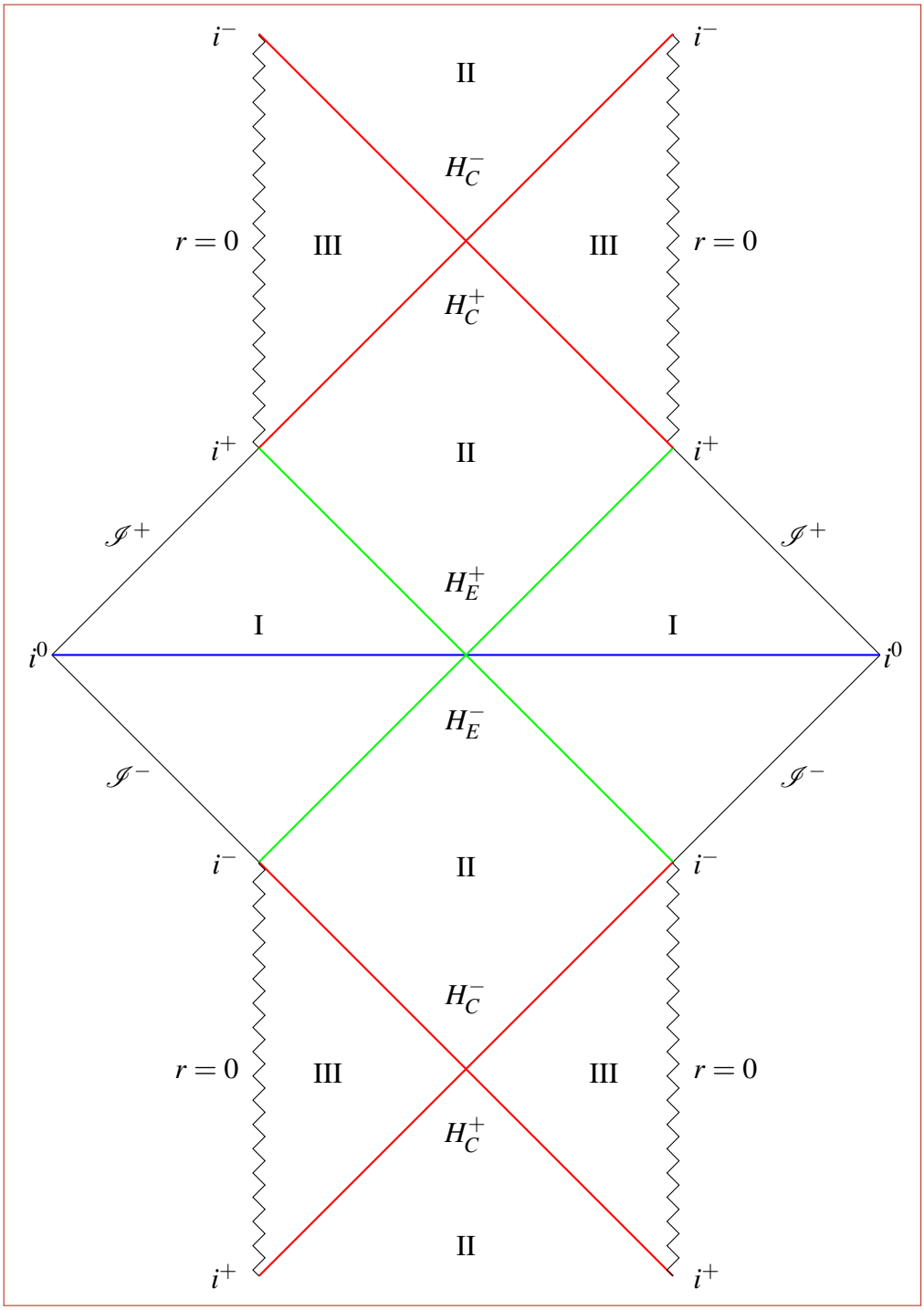
We return to our fd timelike geodesic observer, who (unlike the incoming radial lightlike geodesic just discussed), after having crossed the future Cauchy horizon H_C^+ bounces back from the singularity, increases r, and is even able to cross the next Cauchy horizon H_C^- at $r = r_-$. Strangely enough, this makes him outgoing rather than ingoing. Naively, this would lead him back to the area II where he came from, but according to Theorem 9.2 this is impossible for a fd timelike curve. Therefore, he has entered a new region, interpreted as the interior of a white hole, from which he can move on to cross its "anti" event horizon H_E^- and enter a new asymptotically flat region. The process may then be repeated, which leads to (and in turn is illustrated by) the Penrose diagram on the next page. Adding the south-east diamonds I and II to the original north-west diamonds I and II should be familiar from the Kruskal extension of Schwarzschild space-time, whose extension ends there; recall that regions II are then triangles whose northern or southern borders represents a singularity. Now, however, our ingoing fd timelike observer can cross either of the red $r = r_-$ lines into one of the triangular regions III, and move on to the new region II as described above. To make this space-time geodesically complete also into the past (except at the singularities), one extends the original space-time analogously, to the "south".

If we take the blue horizontal bar as a wannabe Cauchy surface Σ in the extended space-time, the first two red lines above it form its future Cauchy horizon $H^+(\Sigma)$ whilst those to the south of the region II below the blue line form its past Cauchy horizon $H^-(\Sigma)$. Indeed, in the triangular regions III north of this horizon one may initiate inextendible timelike curves that crash into the singularity southward and go on indefinitely northbound. Such curves do not cross Σ; contradicting the definition of a Cauchy surface (and similarly to the past).

A similar tower may be drawn for the extremal case $0 \neq |e| = m$, where compared to the previous case things are simplified by the coalescence $r_+ = r_-$, so that there is just one type of horizon $H^\pm$ that is simultaneously an event horizon $H_E^\pm$ and a Cauchy horizon $H_C^\pm$. Thus one simply places the entire diagram shown on top of and below itself in such a way that the red H^- lines match, and repeats this procedure. A given region III then acts as a black hole towards the lower region I and as a white hole towards the upper one. The difference with the Schwarzschild solution comes from the difference between the functions

$$f(r) = (r-2m)/r; \qquad\qquad h(r) = (r-m)^2/r^2 \tag{9.101}$$

in the metric, which means that in the original coordinates (t, r, θ, φ) neither ∂_t nor ∂_r changes its causal nature if it crosses the horizon. In particular, ∂_r remains spacelike and this makes the singularity timelike (as it is in the two other regimes of the solution).



Penrose diagram of the maximally extended Reissner–Nordström solution for $0 < |e| < m$. *Region I corresponds to* $r > r_+$, *region II to* $r_- < r < r_+$, *and region III to* $0 < r < r_-$. *The repetition is such that a green cross with the associated null infinities* $\mathscr{I}^\pm$ *is added both above and below the red crosses, after which a red cross and the associated* $r = 0$ *singularities are added above and below,* etc. *Compared to the earlier diagram, the new regions make the space-time geodesically complete except at the singularities, and hence it is inextendible. Each green cross is an event horizon (and even a bifurcate Killing horizon, see §10.8), whereas each red cross is a Cauchy horizon with respect to some generic wannabe Cauchy surface, like the one drawn in blue.*[468]

[468] Redrawn from Hawking & Ellis (1973), page 158. The labeling of the regions differs from the Kruskal one.

It is interesting to see some of the differences between Schwarzschild and Reissner–Nordström from the procedure used to merge the solutions (9.44) and (9.45) into a single (Kruskal) space-time.[469] For $0 < |e| < m$, analogously to footnote 445 we use (9.34) - (9.35) and approximate

$$r - r_+ \approx \pm \frac{1}{2\kappa_+} e^{2\kappa_+ r_*} = \pm \frac{1}{2\kappa_+} e^{\kappa(v-u)}, \tag{9.102}$$

where the upper sign applies to $r > r_+$ and the lower one to $r_- < r < r_+$. This gives the approximation $h(r) \approx \pm \exp(\kappa_+(v-u))$. Defining u and v as in (9.34) - (9.35), as $r \approx r_+$, the metric (9.88) is then regularized by the new coordinates

$$U_+ = \mp e^{-\kappa_+ u}; \qquad\qquad V_+ = e^{\kappa_+ v}, \tag{9.103}$$

since $g_{RN} \approx -\kappa_+^{-2} dU_+ dV_+ + \cdots$, as before. More precisely, since (9.56) is now replaced by

$$U_+ V_+ = -e^{2\kappa_+ r_*} = -e^{2\kappa_+ r} (r - r_+)(r - r_-)^{-r_-^2/r_+^2}. \tag{9.104}$$

Together with (9.103) and (9.93), this gives the exact metric in U_+-V_+ coordinates as

$$ds_+^2 = -\frac{e^{-2\kappa_+ r}}{\kappa_+^2 r^2} (r - r_-)^{1+(r_-^2/r_+^2)} dU_+ dV_+ + r^2 d\Omega, \tag{9.105}$$

where $r = r(U_+, V_+)$ is defined via (9.104), i.e. via (9.103), (9.91), and (9.34) - (9.35), as usual.[470] For $r \approx r_-$ we have $r - r_- \approx (U_+ V_+)^{-r_+^2/r_-^2}$, so that $r = r_-$ corresponds to $U_+ V_+ = \infty$ and hence is out of the range of the (U_+, V_+) coordinates. To get to $r = r_-$ and *a fortiori* to $r \to 0$, we introduce new coordinates (U_-, V_-) by making the replacements

$$\kappa_+ \rightsquigarrow \kappa_- = \tfrac{1}{2} h'(r_-) = -\frac{r_+ - r_-}{2r_-^2}; \tag{9.106}$$

$$U_+ \rightsquigarrow U_- = \mp e^{\kappa_- u}; \qquad V_+ \rightsquigarrow V_- = -e^{-\kappa_- v}; \tag{9.107}$$

$$U_+ V_+ \rightsquigarrow U_- V_- = -e^{-2\kappa_- r_*} = -e^{-2\kappa_- r} (r - r_-)(r - r_+)^{-r_+^2/r_-^2}; \tag{9.108}$$

$$ds_-^2 = -\frac{e^{-2\kappa_- r}}{\kappa_-^2 r^2} (r - r_+)^{1+(r_+^2/r_-^2)} dU_- dV_- + r^2 d\Omega, \tag{9.109}$$

where this time the upper sign refers to $r_- < r < r_+$ and the lower one to $0 < r < r_-$. This metric is singular at $r = r_+$, so that unlike the Schwarzschild case, but somewhat like de Sitter, there isn't a single coordinate system that adequately describes the merger. Referring to the above Penrose diagram, the interior of the large diamond consisting of the regions I and II from the original space-time, plus the new regions I and II south-east of those (which totality is similar to the entire Kruskal space-time) is described by the (U_+, V_+) coordinates, which however break down near the border lines $r = r_-$ of the large diamond (both north and south). Unlike the Kruskal case (in which $r_- = 0$) these can be crossed, but this crossing must be described in the new coordinates (U_-, V_-), which can be started in regions II and extend to regions III, *etc.*

However interesting all this may be, similar comments apply as in the Schwarzschild case: realistic collapse is not expected to lead to solutions (and Penrose diagrams) like this, although an exact solution showing this seems lacking.[471] In addition, the interior part of the solution seems unstable; in particular, the Cachy horizon is believed to turn into a curvature singularity under small perturbations, including even such small effects as an observer trying to cross it. This is a major point in favour of Penrose's (strong) cosmic censorship hypothesis; see §10.4.

[469] We here essentially follows Poisson (2004), §5.2.

[470] If $e = 0$, then (9.105) is not quite (9.55); it would be if the constant $-2m\ln(2m)$ in (9.38) were omitted.

[471] See e.g. Sanchis-Gual *et al.* (2016) for some non-rigorous work in this direction.

9.6 The Kerr solution

The last, and physically most relevant, solution to the vacuum Einstein equations we discuss is

$$\begin{aligned} g_K &= -dt^2 + \frac{2mr}{\rho^2}(a\sin^2\theta d\varphi - dt)^2 + \rho^2(\Delta^{-1}dr^2 + d\theta^2) + (r^2+a^2)\sin^2\theta d\varphi^2 \\ &= -\left(1 - \frac{2mr}{\rho^2}\right)dt^2 - \frac{4mar\sin^2\theta}{\rho^2}dtd\varphi + \frac{\rho^2}{\Delta}dr^2 + \rho^2 d\theta^2 + \frac{\Sigma}{\rho^2}\sin^2\theta\, d\varphi^2; \end{aligned} \tag{9.110}$$

$$\Delta := r^2 - 2mr + a^2 = r^2\left(1 - \frac{2m}{r} + \frac{a^2}{r^2}\right); \tag{9.111}$$

$$\rho^2 := r^2 + a^2\cos^2\theta; \tag{9.112}$$

$$\Sigma := (a^2+r^2)^2 - a^2\Delta\sin^2\theta. \tag{9.113}$$

This is the ***Kerr metric***,[472] parametrized by $m > 0$ and $a \in \mathbb{R}\backslash\{0\}$, expressed in ***Boyer–Lindquist coordinates***.[473] These coordinates (t, r, θ, φ) look the same as those in the Schwarzschild solution (9.15), but this analogy is partly misleading and only makes sense for $r > 2m$.[474] In that case, (r, θ, φ) are the usual spherical polar coordinates, and t is the usual time coordinate. In particular, an important difference with both the Schwarzschild and Reissner–Nordström space-times is that the "radial" coordinate r now takes values in $\mathbb{R}$ (as does t). The set $\{(t, r, \theta, \varphi)\}$ where $\theta \in [0, \pi]$ and $\varphi \in [0, 2\pi)\}$ is therefore (topologically) a two-sphere at any fixed (t, r), *even if* $r = 0$. The curvature singularity, which in the Schwarzschild metric is located at $r = 0$ and hence is a *point* at any fixed time t, is now located at $\rho^2 = 0$. At fixed t this set is (topologically) a *circle* (called a *ring* in this context), and the entire set $\rho^2 = 0$, i.e. $r = \cos\theta = 0$, is given by

$$\mathscr{R} := \{(t, r = 0, \theta = \tfrac{1}{2}\pi, \varphi) \mid t \in \mathbb{R}, \varphi \in [0, 2\pi)\}. \tag{9.114}$$

Indeed, the Kretschmann scalar for (9.110) is given by the following generalization of (9.18):

$$R^{\rho\sigma\mu\nu}R_{\rho\sigma\mu\nu} = \frac{48m^2}{\rho^{12}}(r^2 - a^2\cos^2\theta)(\rho^4 - 16a^2r^2\cos^2\theta), \tag{9.115}$$

which blows up in $\mathscr{R}$. Apart from (9.114) there are no other singularities of (9.110) except coordinate issues, and so we take the (preliminary) manifold underlying Kerr space-time to be

$$M = (\mathbb{R}^2 \times S^2)\backslash\mathscr{R}. \tag{9.116}$$

We have not yet found the right coordinates on all of (9.116), since the metric (9.110) looks singular also outside the ring $\mathscr{R}$, namely where $\Delta = 0$. This can be overcome in a similar way as for the Schwarzschild and Reissner–Nordström metrics, namely by passing to Eddington–Finkelstein coordinates. However, before doing so, we can already look at some interesting geodesics. The details depend on a case distinction similar to the one for Reissner–Nordström:

[472] The is metric was discovered by Kerr (1963); see Melia (2009) for the history of this discovery as well as biographical information about Kerr. An exhaustive mathematical treatment of the Kerr metric is given in the monographs by Chandrasekhar (1983) and O'Neill (1995), whereas the volume edited by Wiltshire, Visser, & Scott (2009) is more physics oriented. The introduction to this volume, available in preprint form as Visser (2006) is a nice first introduction, as is Heinicke & Hehl (2015). Among the general GR texbooks, the one by Plebański & Krasiński (2006) also gives very detailed coverage. The Les Houches lectures by Carter (1973) remain valuable.

[473] These coordinates were introduced in Boyer & Lindquist (1967). During a brief visit to the Center for Relativity at the University of Texas, Austin, which is also where Kerr discovered his metric, Robert Boyer (1933–1966) was tragically killed by Charles Whitman in the University Tower shooting massacre on August 1, 1966 (Melia, 2009).

[474] This places (t, r, θ, φ) outside the ergosphere and hence *a fortiori* outside any relevant horizon, see below.

- $0 < |a| < m$, called the ***slowly rotating case*** (comparable with $0 < |e| < m$), which is astrophysically relevant. Then Δ has two distinct zeros, which, as in (9.89), are given by

$$r_\pm = m \pm \sqrt{m^2 - a^2}. \tag{9.117}$$

  It turns out that $r = r_+$ gives the event horizon (as in the Schwarzschild case $a = 0$, where $r_+ = 2m$), but r_- is a Cauchy horizon (as for the Reissner–Nordström metric).

- $0 < |a| = m$, the ***extremal*** case (comparable with $|e| = m$), where $r_+ = r_-$.

- $0 < m < |a|$, the ***rapidly rotating case*** (comparable with $0 < m < |e|$), where $\rho > 0$.

The interpretation of these cases, suggested by their names, comes from the fact that, due to it being stationary, axisymmetric, and asymptotically flat, the Kerr solution has well-defined total mass/energy $\mathscr{E}$ and angular momentum J. These may be defined by the ***Komar formulae***:[475]

$$\mathscr{E} := -\frac{1}{8\pi}\int_{S^2_r} d\sigma_{\mu\nu}\nabla^\mu T^\nu; \qquad \mathscr{J} := \frac{1}{16\pi}\int_{S^2_r} d\sigma_{\mu\nu}\nabla^\mu A^\nu, \tag{9.118}$$

where (at least in the asymptotic region) $T = \partial_t$ is the Killing vector field defining stationarity, $A = \partial_\varphi$ is the Killing vector field defining axial symmetry. The surface element is given by

$$d\sigma_{\mu\nu} = (n_\mu N_\nu - n_\nu N_\mu)\, d^2\sigma, \tag{9.119}$$

where $d^2\sigma$ was defined below (8.103). One takes a spacelike wannabe Cauchy surface $\Sigma \subset M$ (since Kerr space-time is not globally hyperbolic this is all one can do), with fd timelike normal N, containing a sphere S^2_r in the asymptotically flat region, with outward normal n relative to the embedding $S^2_r \hookrightarrow \Sigma$. It can then be shown that $\mathscr{E}$ and $\mathscr{J}$ are independent of Σ and S^2_r, and yield

$$\mathscr{E} = m; \qquad \mathscr{J} = am. \tag{9.120}$$

Thus the metric (9.110) describes a space-time rotating with constant angular velocity. It is stationary but not static: the solution is not invariant under $t \mapsto -t$ but under the double inversion

$$(t, \varphi) \mapsto (-t, -\varphi). \tag{9.121}$$

This is what one would indeed expect of an object rotating with constant angular velocity, where φ is the angle of rotation, since reversing time also reverses the direction of rotation.

We now turn to geodesic motion, starting with a more abstract perspective on the Schwarzschild constants of motion E and L, cf. (9.21). Let X be a Killing vector field, so that $\mathscr{L}_X g = 0$, i.e.,

$$g(\nabla_Y X, Z) + g(\nabla_Z X, Y) = 0 \quad \text{for all } Y, Z \in \mathfrak{X}(M). \tag{9.122}$$

For an observer with four-velocity $u = \dot\gamma$ moving along a causal geodesic γ, eq. (9.122) plus the geodesic equation $\nabla_u u = 0$ make $g(u, X)$ a constant of motion, since taking $Y = Z = u$ in (9.122) gives $g_{\gamma(s)}(\nabla_u X, u) = 0$, and hence, since $\nabla_u u = 0$ because γ is a geodesic,

$$\frac{d}{ds} g_{\gamma(s)}(X, u) = \nabla_u (g_{\gamma(s)}(X, u)) = g_{\gamma(s)}(\nabla_u X, u) + g_{\gamma(s)}(X, \nabla_u u) = 0. \tag{9.123}$$

[475] See e.g. Gourgoulhon (2012), §8.6. The computation of $\mathscr{J}$ was first done by Kerr himself, see Melia (2017), page 75. The computation of $\mathscr{E}$, which coincides with Π^0 in (8.126), is similar to the Schwarzschild case, since one may neglect the a^2/r^2 term in Δ in (9.111), and many other terms drop out by symmetry.

Hence apart from the constant of (geodesic) motion $g(u,u)$, whose value depends on the choice of the affine parameter s and may be fixed to $-\mathsf{m}^2$, where m is the mass of the body moving on the geodesic, our observer carries at least as many conserved quantities as there are linearly independent Killing vector fields. For the Kerr metric this gives

$$E := -g(u,\partial_t); \qquad L := g(u,\partial_\varphi), \tag{9.124}$$

interpreted as its energy and (azimuthal) angular momentum, respectively. If $L=0$, then

$$\frac{d\varphi(t)}{dt} = -\frac{g_{t\varphi}}{g_{\varphi\varphi}} = \frac{2mar}{\Sigma} =: \omega, \tag{9.125}$$

which means that stationary observers rotate with the black hole (***inertial frame dragging***).

Surprisingly, the Kerr metric leads to a fourth constant of motion along geodesics, which is not explicable in terms of isometries of the metric (and remains somewhat mysterious). It was discovered by Carter and may therefore be called C. These four constants of motion turn the four second-order geodesic equations into a first-order system,[476] which for $\mathsf{m}=0$ reads:[477]

$$\Delta\rho\dot{t} = \Sigma E - 2marL; \tag{9.126}$$

$$\rho^2\dot{r}^2 = E^2r^4 + (a^2E^2 - L^2 - C)r^2 + 2m((L-aE)^2 + C)r - a^2C; \tag{9.127}$$

$$\rho^2\dot{\theta}^2 = C + \left(E^2a^2 - \frac{L^2}{\sin^2\theta}\right)\cos^2\theta; \tag{9.128}$$

$$\Delta\rho\dot{\varphi} = 2maEr + (\rho - 2mr)\frac{L}{\sin^2\theta}. \tag{9.129}$$

Compare (9.21); one difference with the Schwarzschild case is that closed geodesic orbits are no longer *necessarily* planar. However, planar orbits do exist and include the Kerr analogues of the unstable photon rings at $r=3m$ in the Schwarzschild metric. These now arise by taking $C=0$ and $\theta=\pi/2$, in which case $\rho^2=r^2$, and (9.127) can be written in a way similar to (9.24), viz.

$$\dot{r}^2 + V(r) = E^2; \qquad V(r) := \frac{L^2 - a^2E^2}{r^2} - \frac{2m(L-aE)^2}{r^3}, \tag{9.130}$$

cf. (9.26). Photon rings by definition have constant r, and, assuming $0 \le |a| \le m$, solving the ensuing equations $V(r)=E$ and $V'(r)=0$ gives two unstable orbits with constant radii

$$\mathsf{r}_\pm = 2m(1 + \cos(\tfrac{2}{3}(\arccos(\pm|a|/m))). \tag{9.131}$$

Depending on the value of $|a|/m$, these fall in the range $m \le \mathsf{r}_- \le 3m \le \mathsf{r}_+ \le 4m$. For $a=0$ the Schwarzschild case $\mathsf{r}_+ = \mathsf{r}_- = 3m$ is recovered. For $a>0$ the smaller orbit is *prograde* (i.e. co-rotating with the black hole), whereas the larger one is *retrogade* (rotating in the opposite direction). For $C>0$ there are other spherical photon orbits off the equatorial plane $\theta = \frac{1}{2}\pi$.

At the opposite end, one has the Kerr version of radial lightlike geodesics, which solve

$$\dot{t} = \frac{r^2+a^2}{\Delta}; \qquad \dot{r} = \pm 1; \qquad \dot{\varphi} = \frac{a}{\Delta}; \qquad \dot{\theta} = 0, \tag{9.132}$$

at radii r where $\Delta(r) \neq 0$; on the two horizons, where $\Delta(r_\pm)=0$, these orbits are rest photons, which solve (9.126) - (9.129) with $E=L=C=0$. As in the Schwarzschild case, these lightlike geodesics rule the event and Cauchy horizons in the sense of Corollary 10.17 below.

[476] See Plebański & Krasiński (2006), §20.6, §20.7 and O'Neill (1995), chapter 4. We also consulted Teo (2003).

[477] Putting $L=0$ in (9.129) does not directly reproduce (9.125), since also E has to be eliminated from (9.129). This constant is a linear combination of $\dot{t}$ and $\dot{\varphi}$, from which $\dot{t}$ must be eliminated from the condition that $L=0$, where L is also a linear combination of $\dot{t}$ and $\dot{\varphi}$. See e.g. eqs. (20.104) - (20.105) in Plebański & Krasiński (2006).

9.7 Inside the Kerr black hole

We now find coordinates in which the zeros of Δ are overcome, starting with the slowly rotating case $0 < |a| < m$. The starting point is once again to introduce a radial tortoise coordinate $r_*(r)$, but in addition we need a new azimuthal angle $\varphi_\pm = \varphi \pm A(r)$, where r_* and A solve

$$\frac{dr_*(r)}{dr} = \frac{r^2+a^2}{\Delta}; \qquad \frac{dA(r)}{dr} = \frac{a}{\Delta}, \tag{9.133}$$

cf. (9.36). With an appropriate boundary condition these equations are solved by

$$r_* = r + \frac{mr_+}{\sqrt{m^2-a^2}} \ln|r-r_+| - \frac{mr_-}{\sqrt{m^2-a^2}} \ln|r-r_-|; \tag{9.134}$$

$$A = \tfrac{1}{2} \frac{a}{\sqrt{m^2-a^2}} \ln\left|\frac{r-r_+}{r-r_-}\right|. \tag{9.135}$$

We pass to lightlike coordinates $u \equiv v_- = t - r^*$ and $v \equiv v_+ = t + r_*$, cf. (9.34) - (9.35). These, in turn, give ingoing and outgoing coordinates, where we relabel $(u,v) \equiv (v_-, v_+)$, i.e.

$$(v, r, \theta, \varphi_+) \equiv (v_+, r, \theta, \varphi_+); \qquad (u, r, \theta, \varphi_-) \equiv (v_-, r, \theta, \varphi_-). \tag{9.136}$$

Similar to (9.44) - (9.45), the Kerr metric (9.110) then becomes

$$\begin{aligned} g_\pm = &-\left(1 - \frac{2mr}{\rho^2}\right) dv_\pm^2 - \frac{4mar\sin^2\theta}{\rho^2} dv_\pm d\varphi_\pm + \rho^2 d\theta^2 \pm 2dv_\pm dr \\ &+ \frac{\Sigma}{\rho^2} \sin^2\theta\, d\varphi_\pm^2 \mp 2a\sin^2\theta d\varphi_\pm dr. \end{aligned} \tag{9.137}$$

This is regular throughout the Kerr space-time (9.116); the coordinate singularities of (9.110) caused by $\Delta = 0$ have now been removed. Explicit computation of the geodesic equations is much more work now than in the Schwarzschild case, but the result is essentially the same (with $\varphi \rightsquigarrow \varphi_\pm$), namely that for constant $C_\pm$ the following formulae define radial lightlike geodesics:

$$(u(s) = v(0), r(s) = s + C_-, \theta(s) = \theta(0), \varphi_- = \varphi_-(0)); \tag{9.138}$$

$$(v(s) = v(0), r(s) = -s + C_+, \theta(s) = \theta(0), \varphi_+ = \varphi_+(0)), \tag{9.139}$$

which are called ***outgoing*** and ***ingoing***, respectively, similar to the blue and the green Schwarzschild lightlike geodesics drawn in the Kruskal diagram in §9.4. For $r > 2m$ one can also see this in Boyer–Lindquist coordinates, where a "radial" lightlike geodesic $\gamma(s) = (t(s), r(s), \theta(s), \varphi(s))$ still has constant θ, but moving φ. In terms of the constant energy $E = -g(\dot\gamma, \partial_t)$, one finds

$$\dot t = \frac{E(r^2+a^2)}{\Delta}; \qquad \dot r = \pm E; \qquad \dot\theta = 0; \qquad \dot\varphi = \frac{aE}{\Delta}, \tag{9.140}$$

with the upper sign $\dot r = +E$ for outgoing geodesics and the lower sign $\dot r = -E$ for incoming ones. Both lightlike geodesics in (9.138) - (9.139) are future directed if we time-orient (M_K, g_K) as in the Schwarzschild case, cf. (9.47), namely by declaring $\underline{L} = -\partial_r$ *in the new coordinates* $(v, r, \theta, \varphi_*)$, which also here is a lightlike vector, to be future directed. For $r > 2m$ this makes ∂_t, *in the original coordinates* (t, r, θ, φ), which is timelike in that region, also future directed, as it should. In the same *original coordinates* the vector $-\partial_r$ is fd timelike in the region $r_- < r < r_+$.

In the region $r < r_-$ things are more involved. Most remarkably, there is a region near the ring where the vector field ∂_φ is timelike,[478] so that one has closed timelike loops! Hence Kerr space-time is acausal, cf. Definition 5.28. Few people are bothered by this, though, since both physicists and mathematicians trust the Kerr solution only up to r_-, beyond which it is supposed to be unstable (see §10.5 and the corresponding comments at the end of §9.5).

If we vary the starting point $(v(0), r(0), \theta(0), \varphi_+(0))$, the ingoing lightlike geodesics (similarly the outgoing ones) form a null congruence, cf. §6.3; the tangent vector field is traditionally called ℓ. In terms of these, the Kerr metric assumes the amazingly simple ***Kerr–Schild form***

$$g_{\mu\nu} = \eta_{\mu\nu} + \frac{2mr}{\rho^2}\ell_\mu \ell_\nu, \tag{9.141}$$

where η is the Minkowski metric in whatever coordinates are used. This shows, in particular, that for $m = 0$ the Kerr metric is the Minkowski metric, which is not quite obvious from (9.110).

In any case, we may now generalize Theorems 9.1 and 9.2:

Theorem 9.3 *Both horizons $H_\pm = \{(v, r, \theta, \varphi) \mid r = r_\pm\}$ (where $\Delta = 0$) are null hypersurfaces, are homeomorphic to $\mathbb{R} \times S^2$, and are one-way membranes towards smaller values of r.*

Proof. The proof of Theorem 9.1 is easily adjusted. First, since r is constant on $H_\pm$, the induced metric $\tilde{g}$ on $H_\pm$ is simply (9.137) without the two terms containing dr, with determinant

$$\det(\tilde{g}) = -\rho^2 \Delta \sin^2\theta. \tag{9.142}$$

This vanishes at $H_\pm$ (defined as the locus where $\Delta = 0$), so that $H_\pm$ are null hypersurfaces. The other proof of this fact works as well: the normal $L_\pm$ to $H_\pm$ is given by

$$L_\pm = 2(\partial_v + \Omega_\pm \partial_{\varphi_*}); \qquad \Omega_\pm := \frac{a}{2mr_\pm} = \frac{a}{r_\pm^2 + a^2}, \tag{9.143}$$

which is lightlike on $H_\pm$ (we omit the general expression for the normal to a hypersurface $r = c$).

To prove the one-way membrane property, instead of (9.51) - (9.52), we now have

$$g(\dot{c}, \dot{c}) < 0 \qquad \Leftrightarrow \qquad \dot{r}(\dot{v} - a\sin^2\theta\,\dot{\varphi}_*) + \tfrac{1}{2}A < 0; \tag{9.144}$$

$$g(\underline{L}, \dot{c}) < 0, \qquad \Leftrightarrow \qquad \dot{v} - a\sin^2\theta\,\dot{\varphi}_* > 0, \tag{9.145}$$

where we have abbreviated a lengthy expression coming from (9.137) by

$$A := -\left(1 - \frac{2mr}{\rho^2}\right)\dot{v}^2 - \frac{4mar\sin^2\theta}{\rho^2}\dot{v}\dot{\varphi}_* + \frac{\Sigma}{\rho^2}\sin^2\theta\,\dot{\varphi}_*^2 + \rho^2\dot{\theta}^2.$$

At both horizons $H_\pm$, this expression A somewhat miraculously takes the positive definite form

$$A_{|H_\pm} = \frac{\sin^2\theta}{\rho^2}(a\dot{v} - 2mr_\pm\dot{\varphi}_*)^2 + \rho^2\dot{\theta}^2, \tag{9.146}$$

which replaces the terms $r^2(\dot{\theta}^2 + \sin^2\theta\,\dot{\varphi}^2)$ in (9.51), at $r = 2m$. Since $A \geq 0$ at $H_\pm$, the argument for the Schwarzschild case still applies and hence for timelike fd curves we must have

$$\dot{r} < 0, \tag{9.147}$$

[478] For $\theta = \frac{1}{2}\pi$ the prefactor of $d\varphi^2$ in (9.110) equals $a^2 + r^2 + 2ma^2/r$, which is negative for small negative r.

at, and therefore also near $H_\pm$. The final step of the proof also applies here, except that the fd "rest" photons for the Kerr metric are characterized by $r = r_\pm$ (and hence $\dot r = 0$), and $\dot\theta = 0$, but

$$\dot\varphi_* = \Omega_\pm \dot v; \qquad\qquad \dot v > a\sin^2\theta\,\dot\varphi_*. \tag{9.148}$$

Hence φ_* cannot be constant, as is also clear from the fact that, as for Schwarzschild, these photons solve $\nabla_{L_\pm} L_\pm = 0$, with $L_\pm$ given by (9.143). Thus they do hover around on S^2. □

The interpretation of the horizons $H_\pm$ is the same as for Reissner–Nordström: $H_+ = H_E^+$ is *the* event horizon, whereas $H_- = H_C^+$ is a Cauchy horizon. The vector field ∂_r also behaves analogously:[479] it is spacelike for $r > r_+$ and $r < r_-$ and timelike for $r_- < r < r_+$. Observers that cross H_+ and subsequently H_- can therefore avoid the singularity (although they cannot return). The singularity is timelike = locally naked, cf. §10.4, but is covered by an event horizon. We return to these horizons in §10.8; the main point will be that the Killing vector field

$$X := \partial_t + \Omega_+ \partial_\varphi; \qquad\qquad \Omega_+ := \omega(r_+) = \frac{2mar_+}{\Sigma} = \frac{a}{r_+^2 + a^2}, \tag{9.149}$$

which is timelike outside H_+, becomes lightlike at H_+, which thence is called a ***Killing horizon***.

A closely related property of a Kerr black hole is its ***ergosphere***.[480] We first define the ***outer ergosurface*** $\mathscr{E}^+$ (also called the ***stationary limit surface***) and ***inner ergosurface*** $\mathscr{E}^-$ by

$$\mathscr{E}^\pm = \{(t,r,\theta,\varphi) \mid r = r_{\mathscr{E}}^\pm(\theta)\}; \qquad r_{\mathscr{E}}^\pm(\theta) := m \pm \sqrt{m^2 - a^2\cos^2\theta}. \tag{9.150}$$

Writing g_{tt} as $-(r - r_{\mathscr{E}}^+)(r - r_{\mathscr{E}}^-)/\rho^2$, we see that $\mathscr{E}^+$ is where ∂_t changes its causal nature:

- ∂_t is *timelike* at $r > r_{\mathscr{E}}^+(\theta)$;
- *lightlike* at $\mathscr{E}^+$;
- *spacelike* within the ***ergosphere***

$$\mathscr{E} = \{(t,r,\theta,\varphi) \mid r_+ < r < r_{\mathscr{E}}^+(\theta)\}; \tag{9.151}$$

- *lightlike* again at $\mathscr{E}^-$;
- *timelike* again for $r < r_{\mathscr{E}}^-(\theta)$.

In the ergosphere a massive particle cannot be at rest (as it can for $r > r_{\mathscr{E}}^+$), but it can still escape. Moreover, in the ergosphere the energy $E = -g(u,\partial_t)$ of a particle with fd four-velocity u can have either sign (whereas for $r > r_{\mathscr{E}}^+(\theta)$ is positive, since u must be fd). This allows the extraction of energy from the black hole via the so-called ***Penrose process***. Here, a particle coming from infinity with necessarily positive energy $E_{\text{as}} > 0$ falls into the ergosphere, decays into a pair, one of positive energy $E_{\text{pos}} > 0$ and one of negative energy $E_{\text{neg}} < 0$, where $E_{\text{pos}} + E_{\text{neg}} = E_{\text{as}}$. If the positive-energy particle subsequently escapes, which is dynamically possible, an amount

$$E_{\text{pos}} - E_{\text{as}} = -E_{\text{neg}} > 0 \tag{9.152}$$

of energy has been extracted from the black hole. This extraction is at the expense of its angular momentum: since $-g(u,X) = E - \Omega_+ L$ and $-g(u,X) > 0$ outside H_+, we have, outside H_+,

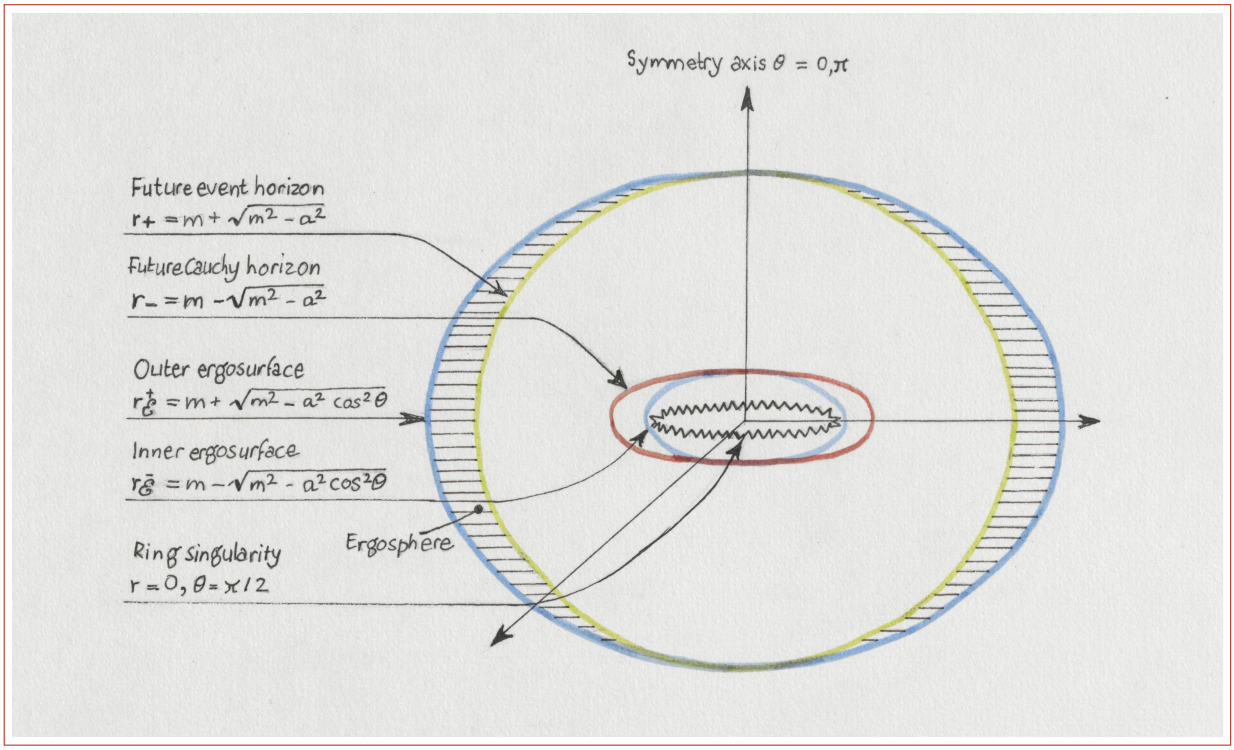
$$E > \Omega_+ L \tag{9.153}$$

where X is the Killing vector field (9.149). Hence $L < E/\Omega_+$, so if the hole absorbs a particle with $E_{\text{neg}} < 0$, it absorbs negative angular momentum $L < E_{\text{neg}}/\Omega_+ < 0$, i.e., loses it. This process could continue until the ergosphere disappear and the black hole stops rotating.

[479] The vector field ∂_θ is spacelike everywhere, whereas ∂_φ is spacelike for $r > 0$, timelike in a certain region near the ring at $r = 0$, where it gives rise to closed timelike loops, and spacelike again for r sufficiently negative.

[480] In Schwarzschild space-time the outer ergosurface coincides with the outer event horizon and hence the ergosphere is empty. The inner event horizon and inner ergosurface both coincide with the (pointlike) singularity.

Here is a picture of the various geometric structures in or near a Kerr black hole.[481]

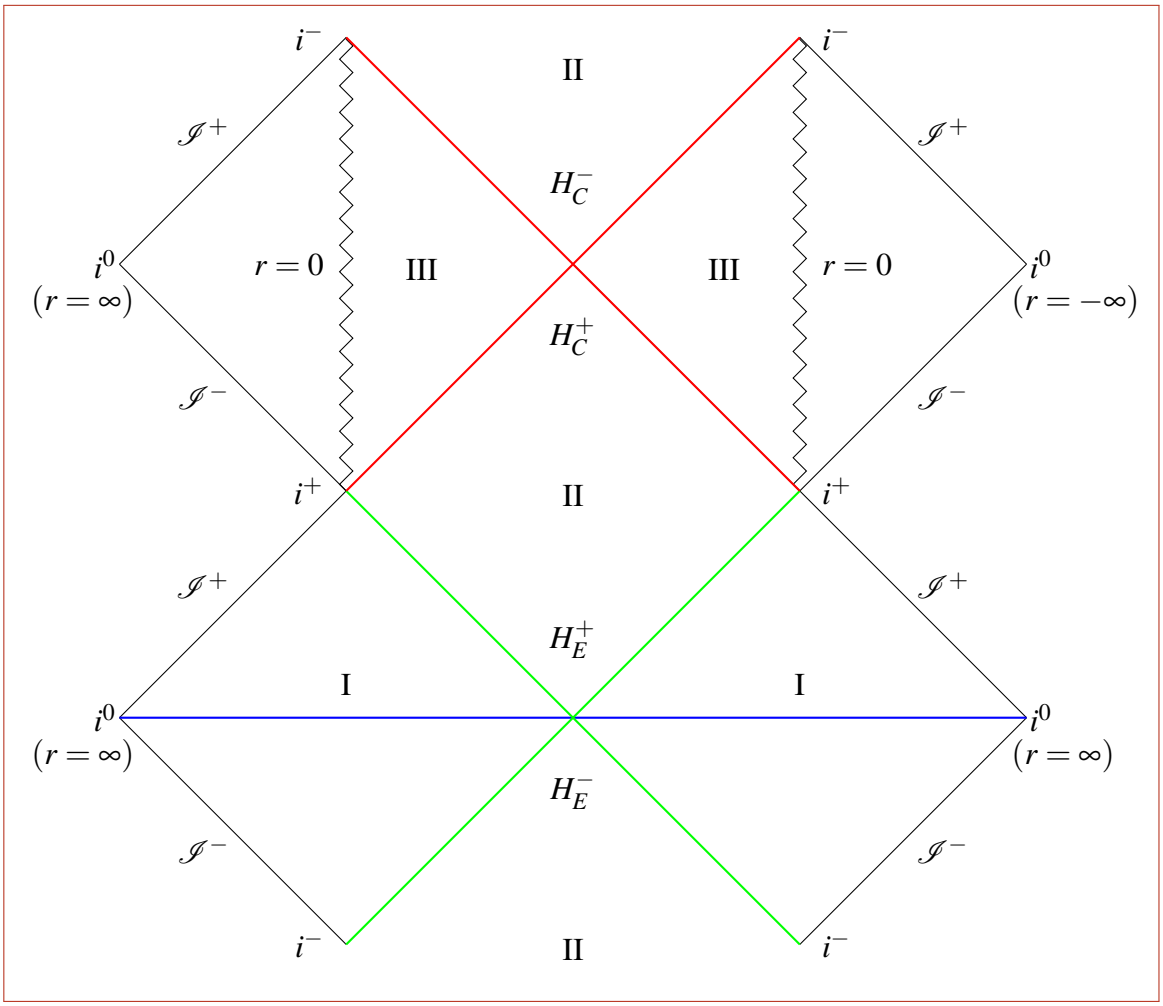


Picture of important $r =$ constant *surfaces in slowly rotating Kerr space-time (in Boyer–Lindquist coordinates) at fixed* t*, shown for* $r \geq 0$ *(although in fact* $r \in \mathbb{R}$*). The event horizons* $H_\pm$ *where* $r = r_\pm$ *are characterized by Theorem 9.3 as one-way membranes:* H_+ *is* the *outer event horizon of a slowly rotating Kerr black hole, since it is the boundary inside which future (null) infinity can no longer be reached. The inner event horizon* H_- *is a Cauchy horizon. Near the singularity* ∂_t *is timelike and* ∂_r *is spacelike, so it can be avoided. The outer ergosphere is the place where the timelike Killing field* ∂_t *switches its causal nature from being timelike at* $r > r^+_{\mathscr{E}}$ *to lightlike at the outer ergosurface, to spacelike until one reaches the inner ergosurface, where it becomes lightlike and then timelike once more. The ergosphere is the region between the outer ergosurface and the outer event horizon; it is the place from which massive particles (or timelike observers) can no longer be at rest, but can still escape to infinity. In the* extremal case *(*$|a| = m > 0$*) both horizons* $H_\pm$ *coalesce, since* $r_+ = r_- = m$*. Furthermore, because* $r^\pm_{\mathscr{E}} = m(1 \pm \sin\theta)$ *the ergosurfaces acquire cusps at* $\theta = 0$ *and* $\theta = \pi$*, at which values all three surfaces touch each other. Otherwise, since* $r^+_{\mathscr{E}} \geq m \geq r^-_{\mathscr{E}}$ *(with equalities iff* $\theta = 0$ *or* $\theta = \pi$*), the single horizon remains enclosed between the outer and inner ergosurfaces. In the* fast case *(*$|a| > m > 0$*) there are no event horizons and the two ergosurfaces have merged into a single (topological) torus with cusps.*[482]

The same misgivings as to the original or maximally extended Schwarzschild solutions apply to this picture as well as to its extensions studied below; notably the instability of the inner event horizon and its extravagant if not crazy causal structure. However, in this case there seems to be no analogue of the exact Oppenheimer–Snyder solution for a rotating black hole.[483]

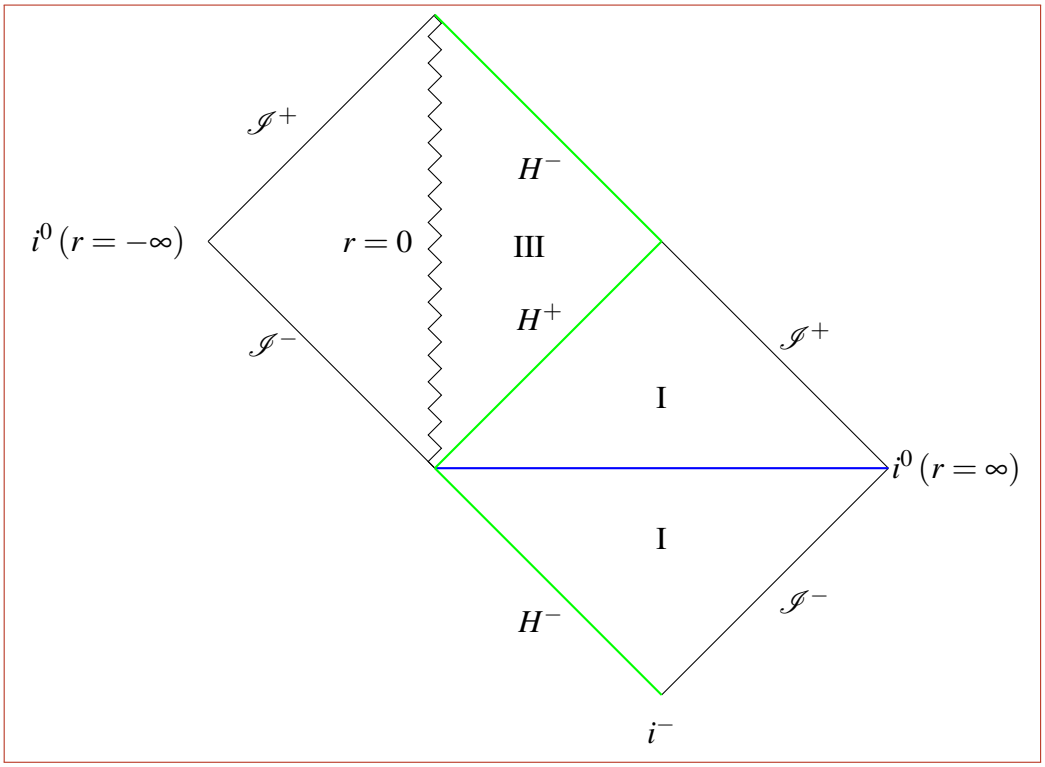
[481] Redrawn from Visser (2006) by Edith de Jong. Explanations and formulae as in the original.
[482] See Carter (1973), §7, and Plebański & Krasiński (2006), §20.5 for pictures of the last two cases.
[483] As a second best see e.g. Nathanail, Most, & Rezzolla (2017) for numerical simulations.

Kerr space-time is geodesically incomplete, and not just at the ring singularity. Hence (except in the fast case, where it is already complete) it can be extended (with all the qualifications and misgivings discussed at the end of §9.5). Here is the relevant Penrose diagram for $0 < |a| < m$, which displays the nature of the (analytic) extension to an inextendible space-time:



Above: *Penrose diagram for the partly extended Kerr solution with $0 < a < m$. The complete extension is an infinite tower: put the part with the green cross on top of the part with the red cross, and put the red part below the green part,* etc. *The range of r is now $(-\infty,\infty)$ instead of $(0,\infty)$, so that (at fixed time) $r = 0$ is a sphere. Penrose diagrams for space-times that lack spherical symmetry (like Kerr) are less effective than for those who are (like Schwarzschild and Reissner–Nordström). In particular, the structure of the (ring) singularity does not come out very well: it is easier for a camel to go through the eye of a needle (i.e. cross the ring singularity) than for a rich man to enter into the kingdom of God.*

Below: *Penrose diagram for the partly extended Kerr solution with $0 < |a| = m$. Region II has now disappeared and event horizons and Cauchy horizons coincide, both simply labeled as $H^\pm = H_E^\pm = H_C^\pm$. This time the infinite tower is built by placing the entire diagram shown on top of and below itself in such a way that the green H^- lines match, and repeating this procedure. In this way region III as shown, which is a black hole for region I shown, becomes a white hole for the new region I NE of the shown region III. Similarly, the new region III SE of the shown region I is a white hole for the latter (the distinction between black and white holes thus fades, or rather depends on which region the hole connects with).*



The maximal (analytic) extensions displayed here can be determined on the basis of the (in)completeness of radial geodesics alone, so that we may freeze θ and φ. Doing this shows that the situation is very similar to Reissner–Nordström: with u and v defined by (9.34) - (9.35), where r_* depends on the particular case, the Reissner–Nordström and Kerr metrics with $\theta = 0$ and either φ_+ (ingoing) or φ_- (outgoing) constant are given by

$$g_{RN} = -h(r)dudv; \qquad h(r) = \frac{(r-r_+^{RN})(r-r_-^{RN})}{r^2}; \qquad r_\pm^{RN} := m \pm \sqrt{m^2-e^2}; \tag{9.154}$$

$$g_K = -k(r)dudv; \qquad k(r) = \frac{(r-r_+^{K})(r-r_-^{K})}{r^2+a^2}; \qquad r_\pm^{K} := m \pm \sqrt{m^2-a^2}, \tag{9.155}$$

respectively; see (9.89), (9.95), (9.137), and (9.117). These horizons look analogous, but (9.154) has a singularity whereas (9.155) does not. Since the singularity in the actual (full) solutions are timelike in both cases and therefore can be avoided, this difference turns out not to matter and the maximal extension of Kerr space-time is essentially the same as for Reissner–Nordström. The only difference lies in the structure of the diamonds III: the role of the singularity at $r = 0$ in Reissner–Nordström is now played by the null infinities $\mathscr{I}^\pm$ at $r = -\infty$.

Apart from satisfying curiosity, the aim of the maximal extension is the following:[484]

Theorem 9.4 *The maximally extended Kerr space-time (M_K^*, g_K^*) for $0 < |a| \leq m$ is geodesically complete, except for geodesics moving into the ring singularity (9.114), which are all incomplete. In particular, (M_K^*, g_K^*) is (smoothly) inextendible, cf. Proposition 6.2.*

For the record, the ***Kerr–Newman metric*** is obtained by changing Δ in the Kerr metric, see (9.111), by $\Delta_e = r^2 - 2mr + a^2 + e^2$. This turns out to be a solution to the Einstein–Maxwell equations with axisymmetric vector potential $A = -er(dt - a\sin^2\theta d\varphi)/\rho^2$, cf. (9.112).

[484] See O'Neill (1995), Theorem 4.3.1, with a 100-page proof through an explicit classification of all geodesics.

10 Black holes II: General theory

The model-independent theory of black holes is based on techniques that were largely developed by Penrose in the 1960s. These techniques were initially motivated by the study of gravitational radiation, but they could also be applied to black holes, as e.g. in the famous paper from 1965 for which Penrose was awarded half of the 2020 Nobel Prize for Physics (see chapter 6).

In his wake, Hawking and others also made important contributions to the abstract study of black holes. Around 1970, this led to a mathematical definition of a black hole and its event horizon, see (10.78) and (10.79) below,[485] which is based on Penrose's idea of *null infinity* and the associated notion of a *conformal completion* of space-time.[486] According to this definition it is not the singularity but the event horizon that defines a black hole. This seems reasonable, since it is the event horizon that makes the hole "black". However, Penrose's 1965 singularity theorem, i.e. Theorem 6.15, does not say anything about event horizons–in fact, the theorem even makes event horizons unnecessary as a means for covering singularities, because the assumption of a Cauchy surface already suffices to make these invisible to the outside world (see especially Corollary 10.10 below). To overcome the discrepancy between his theorem and *black* holes, Penrose launched his great *cosmic censorship conjectures*, which we will discuss in detail.

We then analyze the structure of various black hole horizons, notably event horizons, Cauchy horizons, and Killing horizons, and discuss the uniqueness or "no hair" theorems for black holes. These culminate in Penrose's "final state conjecture" and the associated *Penrose inequality*. We close this chapter with a brief survey of the amazing laws of black hole thermodynamics. Although these laws can be formulated and even proved within classical GR, they can only be understood if quantum (field) theory is invoked. Alas, this exceeds the scope of our book.

---

[485] The following information is slightly adapted from the appendix of Landsman (2021), provided by Eric Curiel. 'Penrose (1968), p. 188, defines an event horizon as the boundary of the chronological past of a timelike curve (essentially the same definition, including the name, as given by Rindler 1956), and notes (p. 206) that $r = 2m$ in Schwarzschild is one. The term "black hole" does not appear in that essay, nor any definition remotely like 'the complement of the causal past of future null infinity'. Penrose (1969), which is one of the most important and visionary papers ever written about gravitational collapse and black holes, does use the term "black hole" (probably the first use in the academic general relativity literature, though the term itself was apparently already used in the early 1960s by Dicke in discussion with a popular science writer), but he always encloses it in scare quotes. In footnote 3 on p. 1146, Penrose almost literally gives the definition (10.79) of an 'absolute event horizon', written in words as 'the boundary of the union of all timelike curves which escape to this external future infinity' and in a formula as $\partial(I^-\mathscr{I}^+)$, [and] he does so in the context of [weakly] asymptotically simple space-times, which [include] black holes! [This] seems to be the first appearance of definition (10.79) in the literature. Carter (1971b) does not give a formal definition of "black hole", but he does give an informal definition of 'domain of outer communication', and says (p. 331) that ' "black holes" [are] regions of space-time beyond the domain of outer communication.' The first explicit *definition* of a "black hole" as the 'connected component of the complement of the causal past of future null infinity' is in Hawking (1972). This is repeated in Hawking & Ellis (1973), §9.2, and seems to have been standard ever since (at least in mathematical physics).' See also footnote 626 for the historical connection with Hawking's area law, which was predicated on using the *absolute* event horizon $\partial I^-(\mathscr{I}^+)$.

[486] When Penrose introduced conformal completions and the ensuing diagrams now named after him in GR (see below), these provided a completely new way of looking at boundary conditions and asymptotic flatness (Friedrich, 2011). Since Penrose started in algebraic geometry (as a PhD student of Hodge in Cambridge, later switching to Todd), in finding both conformal completions and the associated diagrams he was undoubtedly influenced by the theory of Riemann surfaces–one of whose founders was Weyl (1913), the pioneer of the conformal approach to GR! Indeed, Riemann surfaces may equivalently be defined as either one-dimensional complex manifolds, or as two-dimensional Riemannian manifolds *up to conformal equivalence*. The key examples of the Riemann sphere and the Poincaré upper half-plane and disc $\mathbb{D}$ (both actually first found by Beltrami) will be reviewed in the next section. The Poincaré disc $\mathbb{D}$ lies at the basis of the famous *Circle Limit* woodcuts by Escher (nos. I–IV, dating from 1958–1960), see §4.4 for number IV, with which Penrose was well familiar. See also the Introduction.

10.1 Conformal completions of space-time

Penrose's approach to GR is typically based on conformal transformations (cf. §1.9), i.e.

$$\hat{g} = \Omega^2 g; \qquad \hat{g}_{\mu\nu}(x) = \Omega(x)^2 g_{\mu\nu}, \tag{10.1}$$

where initially $\Omega : M \to (0,\infty)$ is strictly positive.[487] The idea is that Ω decrease near "infinity" in such a way that large g-distances become small with respect to $\hat{g}$, with the goal of bringing "infinity" within a finite $\hat{g}$-distance. To make this idea more precise, we first consider the Euclidean plane $(\mathbb{R}^2, g)$, where g is the usual flat metric. In polar coordinates, this reads

$$g = dr^2 + r^2 d\varphi^2. \tag{10.2}$$

Now consider the two-sphere S^2, whose metric in the usual spherical coordinates (θ,φ) reads

$$\hat{g} = d\theta^2 + \sin^2\theta d\varphi^2. \tag{10.3}$$

Define a diffeomorphism $i : \mathbb{R}^2 \to S^2\backslash N$, with $N = (0,0,1)$ i.e. $\theta = 0$, with its inverse, by

$$i(r,\varphi) := (2\arctan(1/r), \varphi); \qquad i^{-1}(\theta,\varphi) = (1/\tan(\theta/2), \varphi); \tag{10.4}$$

the inverse $\imath^{-1} : S^2\backslash N \to \mathbb{R}^2$ is the familiar stereographic projection.[488] The point is that $i : \mathbb{R}^2 \hookrightarrow S^2$ is a ***conformal embedding***, in the sense that as a relation on $\mathbb{R}^2$ we have

$$i^*\hat{g} = (i^*\Omega^2)g, \tag{10.5}$$

a subtle variation of (10.1), where the conformal factor $\Omega : S^2 \to [0,\infty)$, also defined on N, is

$$\Omega(\theta,\varphi) = 2\sin^2(\theta/2). \tag{10.6}$$

This function is strictly positive on $i(\mathbb{R}^2) = S^2\backslash N$ but vanishes at N, as the image under i of all points at infinity (i.e. $r \to \infty$). This property is crucial in keeping $\hat{g}$ finite whilst g measures ever longer distances towards "infinity". We call $(S^2,\hat{g})$ a ***conformal compactification*** of $(\mathbb{R}^2, g)$.

A beautiful example in the same dimension is the ***Poincaré upper half-plane*** $(\mathbb{H}, g)$, i.e.

$$\mathbb{H} = \{x + iy \in \mathbb{C} \mid y > 0\}; \qquad g = \frac{dx^2 + dy^2}{y^2}. \tag{10.7}$$

which is a model of $2d$ hyperbolic geometry, cf. §4.4. It is related to the ***Poincaré disc*** $(\mathbb{D}, \tilde{g})$,

$$\mathbb{D} = \{x + iy \in \mathbb{C} \mid x^2 + y^2 < 1\}; \qquad \tilde{g} = 4\frac{dx^2 + dy^2}{(1 - x^2 - y^2)^2} \tag{10.8}$$

through the isometry $i : \mathbb{H} \to \mathbb{D}$ defined by the ***Cayley transform***

$$i(z) = \frac{z - i}{z + i}; \qquad i^{-1}(\hat{z}) = \frac{\hat{z} + i}{\hat{z} - i}. \tag{10.9}$$

which is also defined on the boundary $\partial\mathbb{H} = \{x + iy \in \mathbb{C} \mid y = 0\}$ and maps this onto $\mathbb{T}\backslash\{1\}$.

[487] In our notation (which differs from many texts) g is the *physical* metric, usually solving the Einstein equations.

[488] Our spherical coordinates (θ,φ) on S^2 are defined by $(x,y,z) = (\sin\theta\cos\varphi, \sin\theta\sin\varphi, \cos\theta)$. In cartesian coordinates on both $\mathbb{R}^2$ and S^2 we have $i(x,y) = (2x, 2y, x^2 + y^2 - 1)/(x^2 + y^2 + 1)$ and $i^{-1}(x,y,z) = (x,y)/(1 - z)$.

The well-known fact that i is an isometry implies that if we now define $(\hat{\mathbb{H}},\hat{g})$ by

$$\hat{\mathbb{H}} := \overline{\mathbb{D}} = \{x+iy\in\mathbb{C}\,|\,x^2+y^2\leq 1\}; \qquad \hat{g} = dx^2+dy^2, \tag{10.10}$$

then (10.5) holds with conformal factor $\Omega : \overline{\mathbb{D}} \to [0,\infty)$, $\Omega(x,y) = 1-x^2-y^2$. Once again, "infinity" for $(\mathbb{H},g)$, consisting of both the x-axis (which because of the factor $1/y^2$ in g is metrically speaking infinitely far from any point in $\mathbb{H}$, in that it takes infinite arc length to get there via a geodesic) and all other points where $r\to\infty$ ($r=\sqrt{x^2+y^2}$), has been brought into the finite realm, which this time consists not of a single point, as in the case of S^2, but of the circle $\partial\mathbb{D} = \mathbb{T} = \{x+iy\in\mathbb{C}\,|\,x^2+y^2=1\}$, where Ω duly vanishes. The single point $1\in\mathbb{T}$ does absorb the entire "$r=\infty$" infinity of $\mathbb{H}$, whereas the remainder $\mathbb{T}\backslash\{1\}$ takes care of the x-axis. Thus the boundary points in $\overline{\mathbb{D}}$ have a somewhat different status in so far as their origin in $\hat{\mathbb{H}}$ is concerned, but from the point of view of the Riemannian manifold (with boundary) $(\overline{\mathbb{D}},\hat{g})$ itself the symmetries of the model guarantee that these distinctions are lost.[489] Of course, more directly one may also start from $(\mathbb{D},\tilde{g})$ and consider $(\overline{\mathbb{D}},\hat{g})$ to be *its* conformal completion.

Penrose magisterially adapted such examples to a space-time context, as follows:[490]

Definition 10.1 *A* **conformal completion** *of a (non-compact) space-time (M,g) is a space-time $(\hat{M},\hat{g})$, where $\hat{M}$ is a manifold with boundary,*[491] *along with an embedding*

$$i : M \hookrightarrow \hat{M}; \qquad i(M) = \mathrm{int}(\hat{M}) := \hat{M}\backslash\partial\hat{M}, \tag{10.11}$$

that is conformal *in that $i^*\hat{g} = i^*\Omega^2 g$ for some smooth positive function $\Omega : \hat{M}\to\mathbb{R}^+$, such that:*

$$\Omega > 0 \text{ on } i(M); \qquad \Omega = 0 \text{ on } \partial\hat{M}; \qquad d\Omega \neq 0 \text{ on } \partial\hat{M}. \tag{10.12}$$

We also require that the boundary $\partial\hat{M}$ consist of null infinity $\mathscr{I}$ *(pronounced "scri"), in that*

$$\partial\hat{M} = \mathscr{I} := \mathscr{I}^+\cup\mathscr{I}^-; \qquad \mathscr{I}^\pm := \partial\hat{M}\cap J^\pm(M), \tag{10.13}$$

where $J^\pm$ is computed in $\hat{M}$. This defines **future null infinity** $\mathscr{I}^+$ *and* **past null infinity** $\mathscr{I}^-$.

In what follows we often tacitly identify M with $i(M)$, so that $\hat{g}$ and g are related by (10.1), understood to hold on $M\equiv i(M)$ only, rather than all of $\hat{M}$; indeed, g is not defined on $\partial\hat{M}$.

This definition does not fix Ω, but the identification of the boundary $\partial\hat{M}$ with null infinity $\mathscr{I}$ and the conditions (10.12) are well served by choosing Ω such that $\Omega(\gamma(s))\sim C/s$ for some constant C as $s\to\infty$, along all complete lightlike geodesics affinely parametrized by s; see §10.2.

[489] Being an isometry, $i:\mathbb{H}\to\mathbb{D}$ maps geodesics of $(\mathbb{H},g)$ to geodesics of $(\mathbb{D},\tilde{g})$. The former are either semicircles hitting the x-axis at straight angles, or straight vertical lines. The latter are segments of circles that intersect $\partial\mathbb{D}$ orthogonally, including the limiting case of straight lines through the origin (which need not be images of straight lines in $\mathbb{H}$, not even when one of the endpoints happens to be $1\in\mathbb{T}$). See e.g. Beardon (1983), chapter 7.

[490] See Penrose (1964), who–in the context of gravitational waves–adds the condition that every lightlike geodesic has two end-points on $\partial\hat{M}$, defining (M,g) to be ***asymptotically simple***. This excludes black hole space-times and so we will not use it, following Chruściel (2020), §3.1. For more on conformal completions see Hawking & Ellis (1973), §6.9, Geroch (1977), Wald (1984), §11.1, Penrose & Rindler (1986), chapter 9, Stewart (1991), chapter 3, Frauendiener (2000), and Valiente Kroon (2016). In connection with asymptotic flatness, see also footnote 497.

[491] See §2.6. Note that $(\hat{M},\hat{g})$ has no corners, unlike Penrose diagrams: points such as i^0 and $i^\pm$ of such diagrams are *not* included in $\hat{M}$, which unlike the Riemannian examples is *not* compact. We assume that the boundary $\partial\hat{M}$ is smooth; whether this is really the case for specific space-times is a subtle issue, taken up in footnote 497. One may define spatial and timelike conformal completions that include these points (Geroch, 1977; Ashtekar, 1980).

10.2 Conformal completion and Penrose diagrams

We first illustrate the idea of a conformal completion for Minkowski space-time $\mathbb{M}$. It is convenient to move from Cartesian coordinates (x^0,x^1,x^2,x^3) to polar ones (t,r,θ,φ), and replace (t,r) by lightlike coordinates $(u,v)\in\mathbb{R}^2$ (i.e. ∂_u and ∂_v are lightlike vectors), defined as

$$u := t-r, \qquad t = \tfrac{1}{2}(v+u); \tag{10.14}$$
$$v := t+r, \qquad r = \tfrac{1}{2}(v-u), \tag{10.15}$$

so that $v\geq u$, and $v=u$ iff $r=0$. In the coordinates (u,v,θ,φ), the Minkowski metric reads

$$\eta = -du\,dv + \tfrac{1}{4}(u-v)^2(d\theta^2+\sin^2\theta d\varphi^2). \tag{10.16}$$

This formula implies that the curves

$$s\mapsto (u(s),v(s),\theta(s),\varphi(s)) = (u_0,v_0+s,\theta,\varphi); \tag{10.17}$$
$$s\mapsto (u(s),v(s),\theta(s),\varphi(s)) = (u_0-s,v_0,\theta,\varphi); \tag{10.18}$$

both defined for $u_0-v_0\leq s<\infty$, are radial lightlike geodesics: eq. (10.17), where u is constant, is future directed (fd) whilst (10.18), where v is constant, is past directed (pd). In line with the (second) comment following Definition 10.1, we now define Ω initially on $\mathbb{M}$ by

$$\Omega(u,v,\theta,\varphi) := (1+u^2)^{-1/2}(1+v^2)^{-1/2}; \tag{10.19}$$
$$\Omega(p,q,\theta,\varphi) = \cos p\cos q, \tag{10.20}$$

where for later use we have also introduced the 'compactifying' coordinates (p,q) defined by

$$p := \arctan v; \qquad v = \tan p, \tag{10.21}$$
$$q := \arctan u; \qquad u = \tan q, \tag{10.22}$$

where $p,q\in(-\frac{1}{2}\pi,\frac{1}{2}\pi)$ and $p\geq q$. This turns the original and rescaled metrics into

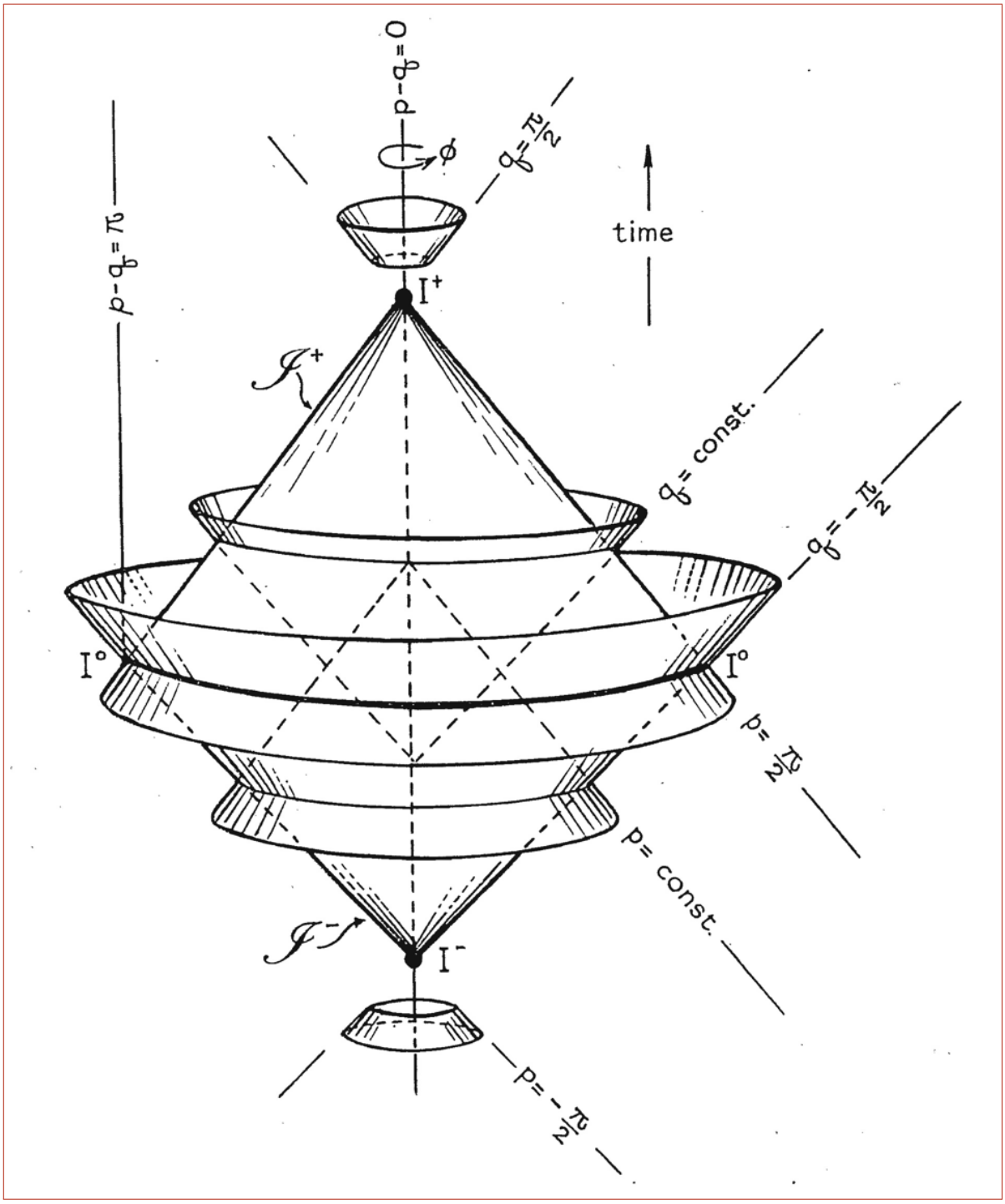
$$\eta = \frac{1}{\cos^2 p\cos^2 q}(-dp\,dq+\tfrac{1}{4}\sin^2(p-q)\cdot(d\theta^2+\sin^2\theta d\varphi^2)); \tag{10.23}$$
$$\hat\eta = \Omega^2\eta = -dp\,dq+\tfrac{1}{4}\sin^2(p-q)\cdot(d\theta^2+\sin^2\theta d\varphi^2). \tag{10.24}$$

We are now in a position to define the conformal completion $(\hat{\mathbb{M}},\hat\eta)$ of $(\mathbb{M},\eta)$ as

$$\hat{\mathbb{M}} := \{(p,q,\theta,\varphi)\mid (p,q)\in(-\tfrac{1}{2}\pi,\tfrac{1}{2}\pi)^2, p\geq q, (\theta,\varphi)\in S^2\}\cup\mathscr{I}^+\cup\mathscr{I}^-; \tag{10.25}$$
$$\mathscr{I}^+ := \{(p,q,\theta,\varphi)\mid p=\tfrac{1}{2}\pi, q\in(-\tfrac{1}{2}\pi,\tfrac{1}{2}\pi),(\theta,\varphi)\in S^2\}; \tag{10.26}$$
$$\mathscr{I}^- := \{(p,q,\theta,\varphi)\mid p\in(-\tfrac{1}{2}\pi,\tfrac{1}{2}\pi), q=-\tfrac{1}{2}\pi,(\theta,\varphi)\in S^2\}, \tag{10.27}$$

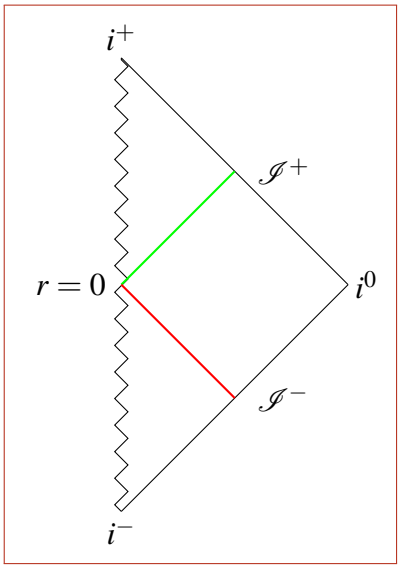
and $\hat\eta$ given by (10.24), now also defined on the boundary $\mathscr{I}=\mathscr{I}^+\cup\mathscr{I}^-$, where it is perfectly regular. Finally, the embedding $i:\mathbb{M}\hookrightarrow\hat{\mathbb{M}}$ is given by $i(u,v,\theta,\varphi)=(\arctan v,\arctan u,\theta,\varphi)$.

A characteristically beautiful drawing of $\hat{\mathbb{M}}$ in Penrose's own hand, including the meaning of the (p,q,φ) coordinates, with the θ-coordinate suppressed, may be found on the next page.



The conformal completion $(\hat{\mathbb{M}}, \hat{\eta})$ *of Minkowski space-time* $(\mathbb{M}, \eta)$*, with the* θ*-coordinate suppressed, taken from Penrose (1964). Future timelike infinity* I^+*, past timelike infinity* I^-*, and spacelike infinity* I^0 *(called* i^+*,* i^-*, and* i^0 *in the main text, following current notation), are drawn, but do not belong to* $\hat{\mathbb{M}}$*. Likewise, the shells at* $p > \frac{1}{2}\pi$ *and* $q < -\frac{1}{2}\pi$ *are not part of* $\hat{\mathbb{M}}$*, which "ends" at* I^0 *and is a rotated diamond without the equatorial circle and north and south poles; they are just drawn to clarify the meaning of the coordinates. Also, the caps above* I^+ *and below* I^- *are not part of* $\hat{\mathbb{M}}$*. Metrically* I^0 *is a point, like* I^+ *and* I^-*, rather than a circle.*

In a ***Penrose diagram*** of $(\mathbb{M},\eta)$, or indeed of any space-time (M,g) admitting a conformal completion, one suppresses the angles (θ,φ) and draws $\hat{\mathbb{M}}/S^2$ (or $\hat{M}/S^2$ in such a way that lightlike geodesics are at $\pm 45^\circ$, as in $\mathbb{M}$ (this leads to some distortions in case g is not spherically symmetric, as e.g. the Kerr metric). This is an important tool for visualizing especially black holes. The points $i^\pm$ and i^0 defined below are typically included in such diagrams, although they are not part of $\hat{M}$. The Penrose diagram of Minkowski space-time, then, is as follows:



Penrose diagram for Minkowski space-time in the (p,q) coordinates, where $(p,q)\in[-\pi/2,\pi/2]$ subject to $p\geq q$; the zigzag line $r=0$ corresponds to $p=q$; it is a boundary to the diagram and as such "singular", but this is an unfortunate coordinate singularity. The green line (constant q) is a fd lightlike geodesic and the red line (constant p) is a pd lightlike geodesic. The three corners are

$$i^- = (-\pi/2,-\pi/2); \qquad i^0 = (\pi/2,-\pi/2); \qquad i^+ = (\pi/2,\pi/2), \tag{10.28}$$

whereas the smooth boundary components are given by, cf. (10.26) - (10.27)*,*

$$\mathscr{I}^+ = \{(p,q) \mid p = \tfrac{1}{2}\pi, q\in(-\tfrac{1}{2}\pi,\tfrac{1}{2}\pi)\}; \tag{10.29}$$
$$\mathscr{I}^- = \{(p,q) \mid q = -\tfrac{1}{2}\pi, p\in(-\tfrac{1}{2}\pi,\tfrac{1}{2}\pi)\}. \tag{10.30}$$

- **Future null infinity** $\mathscr{I}^+$ *corresponds to $v=\infty$ at finite u, i.e. $r\to\infty$ and $t\to\infty$ at fixed $t-r$. All future inextendible fd lightlike geodesics end in $\mathscr{I}^+$, and all its points occur in this way.*

- **Past null infinity** $\mathscr{I}^-$ *corresponds to $u=-\infty$ at finite v, or $r\to\infty$ and $t\to-\infty$ at fixed $t+r$. All past inextendible pd lightlike geodesics end in $\mathscr{I}^-$, and all its points occur in this way.*

- **Future timelike infinity** i^+ *corresponds to $u=v=\infty$, i.e. $t\to\infty$ at finite r, and as such is the single endpoint of all future inextendible fd timelike geodesics.*

- **Past timelike infinity** i^- *corresponds to $u=v=-\infty$, i.e. $t\to-\infty$ at finite r, and is the single endpoint of all past inextendible pd timelike geodesics.*

- **Spacelike infinity** i^0 *corresponds to $u=-\infty$ and $v=\infty$, i.e. $r\to\infty$ at finite t, and is the single endpoint of all inextendible spacelike geodesics.*

Since affinely parametrized radial lightlike geodesics with respect to $\hat{\eta}$ are simply given by

$$\hat{s} \mapsto (p(\hat{s}), q(\hat{s})) = (p_0 + \hat{s}, q_0) \qquad \text{(future directed);} \tag{10.31}$$

$$\hat{s} \mapsto (p(\hat{s}), q(\hat{s})) = (p_0, q_0 - \hat{s}) \qquad \text{(past directed),} \tag{10.32}$$

it should be clear that all points in $\mathscr{I}^+$ and $\mathscr{I}^-$, respectively, are reached by such geodesics, so that the last condition of Definition 10.1 is met. Note that for the geodesic (10.31) we have

$$\Omega((p(\hat{s}), q(\hat{s})) = \cos(p_0 + \hat{s})\cos(q_0) \sim -\cos(q_0)(\hat{s} + p_0 - \tfrac{1}{2}\pi), \tag{10.33}$$

as the $\hat{\eta}$-geodesic $\hat{\gamma}$ in question approaches $\mathscr{I}^+$, i.e. as $p(\hat{s}) \to \frac{1}{2}\pi$. By an affine reparametrization such that $\hat{s} = 0$ when $\hat{\gamma}(\hat{s}) \in \mathscr{I}^+$, we may therefore achieve that near $\mathscr{I}^+$ we have

$$\Omega(\hat{\gamma}(\hat{s})) \sim -\hat{s}. \tag{10.34}$$

From the point of view of the original metric η, by (10.17) and (10.19) the same geodesic, but now affinely parametrized with respect to η and relabeled $\gamma(s)$, i.e. $\gamma(s(\hat{s})) = \hat{\gamma}(\hat{s})$, gives

$$\Omega(\gamma(s)) \sim 1/s, \tag{10.35}$$

as $\mathscr{I}^+$ is approached, i.e. as $s \to \infty$. The reconfirms the name "future null *infinity*" for $\mathscr{I}^+$.

More generally, suppose $\hat{g} = \Omega^2 g$ are conformally related metrics (not necessarily flat or even Lorentzian). Let $\hat{s} \mapsto \hat{\gamma}(\hat{s})$ be a $\hat{g}$-geodesic (which by convention is affinely parametrized), with corresponding g-geodesic $s \mapsto \gamma(s)$. Then a straightforward calculation gives

$$\frac{ds}{d\hat{s}} = \frac{1}{\Omega(\hat{s})^2}. \tag{10.36}$$

Hence if (10.34) holds, as can always be achieved because of (10.12), then (10.35) follows.

We now state some important properties of $\mathscr{I}^\pm$. Since each point in the diagram (excluding $i^\pm$ and i^0) is a two-sphere S^2, eqs. (10.29) - (10.30) give topologically and diffeomorphically,

$$\mathscr{I}^+ \cong \mathscr{I}^- \cong \mathbb{R} \times S^2. \tag{10.37}$$

However, it follows from (10.24) that the would-be two-spheres at $i^\pm$ and i^0 have zero radius and hence should be seen as points (once again, these do not belong to $\hat{\mathbb{M}}$). The wiggly line marked $r = 0$ is indeed a line (i.e. it is homeomorphic to $\mathbb{R}$); its singular appearance as a boundary in the Penrose diagram is a consequence of the fact that such diagrams are pictures of $\hat{M}/SO(3)$ rather than of $\hat{M}$ itself.[492] This can already be seen in the usual (defining) action of $SO(3)$ on $\mathbb{R}^3$, where the quotient $\mathbb{R}^3/SO(3) \cong [0,\infty)$ has zero as a boundary point, corresponding to the fact that the stabilizer of (r,θ,φ) suddenly changes from $SO(2)$ for any $r > 0$ to $SO(3)$ at $r = 0$.

Let us also note that $\mathscr{I}^+$ and $\mathscr{I}^-$ are *null hypersurfaces* (see Definition 4.15); for example, we see from (10.29) that $T_x\mathscr{I}^+$ is spanned by vectors $(\partial/\partial q, \partial/\partial\theta, \partial/\partial\varphi)$, upon which (10.24) shows that $\partial/\partial q$ is both normal and tangent to $\mathscr{I}^+$, and hence lightlike (likewise for $\mathscr{I}^-$ with $q \rightsquigarrow p$). This implies that the metric $\hat{\eta}$ is degenerate on $\mathscr{I}^\pm$; e.g. at future null infinity $\mathscr{I}^+$ it is

$$\hat{\eta}_{|\mathscr{I}^+} = \tfrac{1}{4}(\cos^2 q) \cdot (d\theta^2 + \sin^2\theta d\varphi^2). \tag{10.38}$$

[492] Although a general metric $g \in T^{(2,0)}M$, i.e. "$g_{\mu\nu}$", does not push forward to $M/SO(3)$ under the canonical projection $\pi : M \to M/SO(3)$, its inverse $g^{-1} \in T^{(0,2)}M$, i.e. "$g^{\mu\nu}$", does. As long as the $SO(3)$-orbits are spacelike two-spheres (i.e. even if rotations fail to be isometries), the pushforward $\pi_* g^{-1} \in T^{(0,2)}(M/SO(3))$ is invertible and its inverse $g_2 = (\pi_* g^{-1})^{-1} \in T^{(2,0)}(M/SO(3))$ is a Lorentzian metric of signature $(-+)$ on $M/SO(3)$.

Future (and past) null infinity of $\mathbb{M}$ has another desirable property, which is shared by the usual black hole space-times like Schwarzschild and Kerr, namely *completeness*. It takes some effort to define what this means, but the idea is that at least sufficiently far away (from the black hole, if any), there is no end to the future. This should be expressed technically by the fact that lightlike geodesics within $\mathscr{I}^+$ extend infinitely into the future, but unfortunately this is not the case for the choice of the conformal completion $(\hat{\mathbb{M}}, \hat{\eta}, \Omega)$ used so far:[493] lightlike geodesics within $\mathscr{I}^+$ at constant (θ, φ), and $p = \frac{1}{2}\pi$, simply take the form $q(s) = q_0 + s$ with affine parameter s, and then come to a stop as $q(s) \to \frac{1}{2}\pi$, which of course happens for some $s < \infty$.

This can be remedied by a different choice of Ω and hence of $\hat{\eta}$ (since $\eta = \Omega^2\hat{\eta}$ is fixed), keeping $\hat{\mathbb{M}}$ as it is. We give a systematic formulation in the next section, but for the moment we note that changing Ω to $\Omega' = \omega\Omega$, with $\omega(p,q) = 1/\sin(p-q)$, rescales (10.24) into

$$\hat{\eta}' = -4\frac{dp\,dq}{\sin^2(p-q)} + d\theta^2 + \sin^2\theta\, d\varphi^2. \tag{10.39}$$

It follows that for this metric at $\mathscr{I}^+$, i.e. for $p = \frac{1}{2}\pi$, we have $\Gamma^q_{qq} = -\tan q$, so that lightlike geodesics γ within $\mathscr{I}^+$ at constant (θ, φ) are given, perhaps after affine reparametrization,[494] by

$$\gamma(s) = (p(s), q(s), \theta(s), \varphi(s)) = (\tfrac{1}{2}\pi, \arcsin s, \theta_0, \varphi_0). \tag{10.40}$$

These geodesics rule the null hypersurface $\mathscr{I}^+$ (in that each point of $\mathscr{I}^+$ lies on one of them) and are *complete* (in the usual sense of being defined for all $s \in \mathbb{R}$), reaching the boundary point i^+ of $\mathscr{I}^+$, i.e. future timelike infinity, as $s \to \infty$, and i^0, i.e. spacelike infinity, as $s \to -\infty$.

Finally, for a new perspective on the conformal completion of $\mathbb{M}$ we take the $4d$ cylinder

$$E = \mathbb{R} \times S^3, \tag{10.41}$$

where the 3-sphere $S^3 = \{(x_1, x_2, x_3, x_4) \in \mathbb{R}^4 \mid x_1^2 + x_2^2 + x_3^2 + x_4^2 = 1\} \subset \mathbb{R}^4$ is coordinatized by

$$\begin{aligned} x_1 &= \cos\chi; & x_2 &= \sin\chi\cos\theta; \\ x_3 &= \sin\chi\sin\theta\cos\varphi; & x_4 &= \sin\chi\sin\theta\sin\varphi, \end{aligned} \tag{10.42}$$

where $\chi \in [0, \pi]$, $\theta \in [0, \pi]$, and $\varphi \in [0, 2\pi]$. This space has a Lorentzian metric,[495] given by

$$\hat{g} = -d\tau^2 + g_{S^3} = -d\tau^2 + d\chi^2 + \sin^2\chi \cdot (d\theta^2 + \sin^2\theta\, d\varphi^2). \tag{10.43}$$

To relate this to Minkowski space-time $(\mathbb{M}, \eta)$, recall (10.21) - (10.22) and put

$$\tau = p + q; \qquad \chi = p - q. \tag{10.44}$$

Given the range $p, q \in (-\frac{1}{2}\pi, \frac{1}{2}\pi)$ and $p \geq q$, this yields $\tau \in (-\pi, \pi)$ and $\chi \in (0, \pi)$, so that we may embed $\mathbb{M}$ into E via $i : \mathbb{M} \hookrightarrow \hat{\mathbb{M}}$ as defined before and subsequently regarding $\hat{\mathbb{M}}$ as a subspace of E; the closure of $i(\mathbb{M}) \subset E$ in E is $\hat{M}$ with the corners $i^\pm$ and i^0 added. The embedding $i : \mathbb{M} \hookrightarrow E$ is conformal, since from (10.16) and (10.43) we find the relation

$$\hat{g}_{|\mathbb{M}} = \Omega^2 g, \tag{10.45}$$

where $\Omega : \mathbb{M} \to \mathbb{R}^+$ is given by (10.20). In conclusion, $(\hat{\mathbb{M}}, \hat{\eta})$ is precisely the conformal completion of $(\mathbb{M}, \eta)$ studied above, now embedded in the larger Lorentzian manifold $(E, \hat{g})$.

[493] The next remark follows from the fact that $\Gamma^q_{qq} = 0$ for the metric (10.24), so that $d^2q/ds^2 = 0$.

[494] The general solution of $d^2q/ds^2 = (\tan q) \cdot (dq/ds)^2$ is $q(s) = \arcsin(c_1(c_2 + s))$, for constants c_1, c_2.

[495] Historically, this space-time arose as ***Einstein's static universe***, which is a solution to the Einstein equations with cosmological constant, which indeed Einstein (1917b) introduced precisely to make the universe static.

Null infinity $\mathscr{I}$ for Minkowski space-time $(\mathbb{M},\eta)$ is null, as the name suggests. However, this is not a consequence of the definition: $\mathscr{I}$ can equally well be spacelike or timelike. These possibilities are realized, for example, in the other two Lorentzian manifolds of constant positive and negative curvature, viz. de Sitter space and anti-de Sitter space, respectively (see §4.4). In this light, $(\mathbb{M},\eta)$ has constant curvature and cosmological constant both equal to zero.[496]

We start with de Sitter space dS^4_ρ, defined by (4.92) with parameter ρ; it satisfies the Einstein equations $R_{\mu\nu}=\lambda g_{\mu\nu}$ with cosmological constant $\lambda=3/\rho^2>0$, as follows from (4.85) with $k=1/\rho^2$. For simplicity we set $\rho=1$ and coordinatize $dS^4_1\cong\mathbb{R}\times S^3$ using $(\tau,\chi,\theta,\varphi)$, where $\tau\in\mathbb{R}$ and $(\chi,\theta,\varphi)\in[0,\pi]\times[0,\pi]\times[0,2\pi)$ cover the S^3 part. Specifically, we have

$$x_0=\sinh\tau;\qquad x_1=\cosh\tau\cos\chi;\qquad x_2=\cosh\tau\sin\chi\cos\theta; \tag{10.46}$$

$$x_3=\cosh\tau\sin\chi\sin\theta\cos\varphi;\qquad x_4=\cosh\tau\sin\chi\sin\theta\sin\varphi; \tag{10.47}$$

$$g_+=-d\tau^2+\cosh^2\tau\cdot g_{S^3};\qquad g_{S^3}=d\chi^2+\sin^2\chi\,d\Omega. \tag{10.48}$$

We then compactify $\mathbb{R}$ by switching to $\eta=\arcsin(1/\cosh\tau)=2\arctan(\exp\tau)\in(0,\pi)$, so that

$$g_+=(\sin^{-2}\eta)\cdot(-d\eta^2+g_{S^3}). \tag{10.49}$$

The conformal factor $\Omega(\eta)=\sin\eta$ then turns g into $\hat{g}_+=\Omega^2 g$ already given by (10.43). We see that also dS^4_1 can be conformally embedded into the Einstein universe (10.41), which was in fact how it was discovered. In the absence of spacelike or timelike infinity, a conformal completion of (dS^4_1,g_+) is given by the closure of this image, which simply extends the range of η to $[0,\pi]$. The boundary value $\eta=0$ gives past null infinity $\mathscr{I}^-$, whereas $\eta=\pi$ yields $\mathscr{I}^+$.

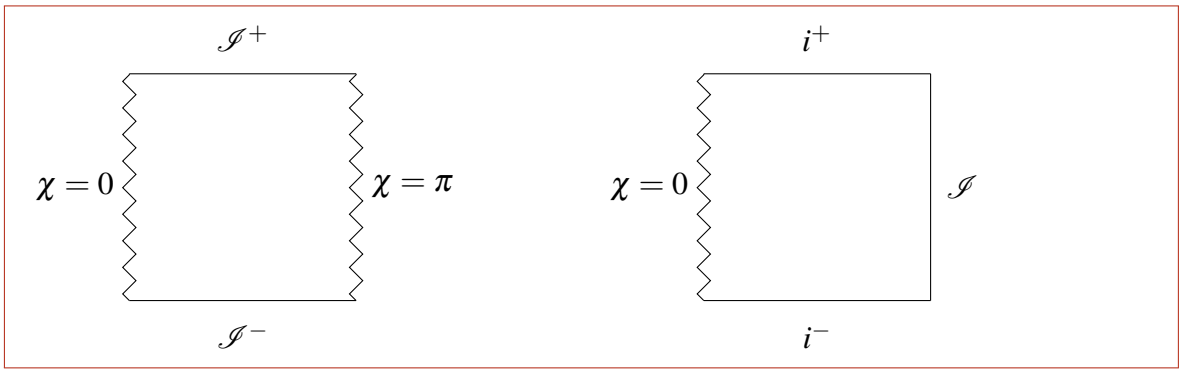
For anti-de Sitter space AdS^4_ρ, cf. (4.93), we have $\lambda=-3/\rho^2<0$, in which we again take $\rho=1$. We use coordinates $\tau\in\mathbb{R}$, $r\geq 0$ or $\chi=\arctan(\sinh r)\in[0,\frac{1}{2}\pi)$, and $(\theta,\varphi)\in S^2$, with

$$x_1=\sinh r\cos\theta;\qquad x_2=\sinh r\sin\theta\cos\varphi;\qquad x_3=\sinh r\sin\theta\sin\varphi; \tag{10.50}$$

$$x_{-1}=\cosh r\cos\tau;\qquad x_0=\cosh r\sin\tau; \tag{10.51}$$

$$g_-=-\cosh^2 r\,d\tau^2+dr^2+\sinh^2 r\cdot g_{S^2}=(\cos^{-2}\chi)\cdot(-d\tau^2+g_{S^3}), \tag{10.52}$$

so that $\Omega(\chi)=\cos\chi$ also conformally embeds (AdS^4_1,g_-) into the Einstein universe. This time, null infinity is connected and timelike, corresponding to $\chi=\frac{1}{2}\pi$, i.e. $r=\infty$. See also §5.10.



Penrose diagrams for de Sitter space (left), where null infinity $\mathscr{I}=\mathscr{I}^+\cup\mathscr{I}^-$ is spacelike and disconnected (and $\chi=0,\pi$ are mere coordinate singularities), and anti-de Sitter space (right), where null infinity $\mathscr{I}$ at $\chi=\frac{1}{2}\pi$ is timelike and connected (and $\chi=0$ is a coordinate singularity; the vertical timelike direction has not been compactified and goes on forever).

[496] See also Griffiths & Podolský (2009), §4.2, 5.2, and Valiente Kroon (2016), §6.3, 6.4 for further details.

10.3 Asymptotic flatness at null infinity and black holes

The example of Minkowski space-time, as well as the black hole space-times reviewed below, suggests the following sharpening of Definition 10.1, which captures both kinds of examples.[497]

Definition 10.2 *A space-time (M,g) is* **asymptotically flat at null infinity** *if it has a conformal completion $(\hat{M},\hat{g})$ with the following additional properties:*[498]

1. *$\mathscr{I}^+ \cong \mathscr{I}^- \cong \mathbb{R}\times S^2$ diffeomorphically, cf. (10.37).*
2. *The Ricci tensor $R_{\mu\nu}$ of the original metric g is such that $R_{\mu\nu} = O(\Omega^3)$ towards $\partial\hat{M}$.*
3. *The lightlike geodesics ruling $\mathscr{I}^\pm$ (which by the previous clause is a null hypersurface in $\hat{M}$) are complete, provided the conformal factor has been chosen such that on $\mathscr{I}^\pm$ one has*

$$\hat{\Delta}\,\Omega = 0. \tag{10.53}$$

In clause 2 and in what follows we tacitly identify M with $i(M)$. Asking $O(\Omega^3)$ is on the safe side (one could use $O(\Omega^{2+\varepsilon})$ for $1/2 < \varepsilon \leq 1$), and implies that $\Omega^{-2}R_{\mu\nu}$ extends by continuity from $i(M)$ to zero on $\partial\hat{M}$, as in $R_{\mu\nu}(r) \sim 1/r^3$ as $r\to\infty$. The simplest way to satisfy this is to assume that (M,g) solves the vacuum Einstein equations $R_{\mu\nu}=0$; in the presence of matter one equivalently asks that $T_{\mu\nu}$ is $O(\Omega^3)$. The third clause, in which $\hat{\Delta} := \hat{g}^{\mu\nu}\hat{\nabla}_\mu\hat{\nabla}_\nu$, makes sense because of a crucial fact, noted (*mutatis mutandis*) without proof in Penrose (1964, 1968):[499]

Proposition 10.3 *On the boundary $\partial\hat{M}$ the scaling function Ω satisfies the eikonal equation*

$$\hat{g}(\hat{\nabla}\Omega,\hat{\nabla}\Omega) = 0, \tag{10.54}$$

so that $\partial\hat{M}$ (more precisely: each connected component thereof) is a null hypersurface in $\hat{M}$.

Proof. A simple computation yields the effect of a conformal rescaling on the Ricci tensor:[500]

$$R_{\mu\nu} = \hat{R}_{\mu\nu} + \Omega^{-1}(2\hat{\nabla}_\mu\hat{\nabla}_\nu\Omega + \hat{g}_{\mu\nu}\hat{\Delta}\,\Omega) - 3\Omega^{-2}\hat{g}(\hat{\nabla}\Omega,\hat{\nabla}\Omega)\hat{g}_{\mu\nu}. \tag{10.55}$$

$$\Rightarrow\ \hat{g}(\hat{\nabla}\Omega,\hat{\nabla}\Omega) = \tfrac{1}{12}(\Omega^2\hat{R} - R) + \tfrac{1}{2}\Omega\,\hat{\Delta}\,\Omega. \tag{10.56}$$

Now $R_{\mu\nu} = O(\Omega^3)$ gives $R = O(\Omega^5)$. Eq. (10.54) follows, as $\hat{g}$ is smooth and $\Omega = 0$ on $\partial\hat{M}$. □

[497] Penrose himself already realized that definitions of this kind, which combine smoothness of the boundary $\partial\hat{M}$ with specific conditions at infinity, imply detailed fall-off (or 'peeling') properties of the Weyl tensor at infinity, which may not always hold. See e.g. Klainerman & Nicolò (2003), Friedrich (2004, 2018), Adamo, Newman, & Kozameh (2012), and Dafermos (2012). For the usual black hole solutions (i.e. Schwarzschild, Reissner–Nordström and Kerr) the boundary is smooth, and this is true much more generally, e.g. for stationary space-times satisfying standard energy conditions (Chruściel et al., 2001). It holds even generically in a suitable topological sense (Chruściel & Delay, 2002; Corvino, 2007; Paetz, 2014; Chruściel & Paetz, 2015), so we will not worry about this.

[498] Clause 3 is due to Geroch & Horowitz (1978). See also Horowitz (1979) and Wald (1984), §11.1.

[499] If $R_{\mu\nu} = \lambda g_{\mu\nu}$, then $\hat{g}(\hat{\nabla}\Omega,\hat{\nabla}\Omega) = -\frac{1}{3}\lambda$ on $\partial\hat{M}$, so that $\hat{\nabla}\Omega$ is timelike and hence $\partial\hat{M}$ is spacelike if $\lambda > 0$, and *vice versa* if $\lambda < 0$ (Penrose, 1964, Lecture II; Penrose, 1968, p. 181). See Ashtekar, Bonga, & Kesavan (2015) and Ashtekar & Magnon (1984), respectively, for theory, and §10.2 above for the two simplest examples. But as the indisputable king of null geometry in GR, Penrose must have taken special pleasure in the case $\lambda = 0$!

[500] It is easily verified by direct computation (see e.g. Valiente Kroon, 2016, §5.2.2; Chruściel, 2020, Appendix H.6) that if $g' = \varphi^2 g$, then $R'_{\mu\nu} = R_{\mu\nu} - \varphi^{-1}(2\nabla_\mu\nabla_\nu\varphi + g_{\mu\nu}\Delta_g\varphi) + \varphi^{-2}(4\nabla_\mu\varphi\nabla_\nu\varphi - g_{\mu\nu}g(\nabla\varphi,\nabla\varphi))$. Now replace $g' \rightsquigarrow g$ and $g \rightsquigarrow \hat{g}$, so that $\varphi = 1/\Omega$. This gives (10.55), which is eq. (11.1.16) in Wald (1984).

Without clause 3, the definition (10.78) below of a black hole would be flawed (see footnote 506). The need for a condition on Ω in order to state completeness of null infinity has been explained in the previous section; for otherwise even $(\hat{\mathbb{M}}, \hat{\eta})$ would be a counterexample. We write

$$\hat{N} := \hat{\nabla}\Omega; \qquad \hat{N}_\mu = \partial_\mu \Omega; \qquad \hat{N}^\mu = \hat{g}^{\mu\nu}\partial_\mu \Omega, \tag{10.57}$$

so that $\hat{N}^\mu \hat{N}_\mu = 0$ on $\mathscr{I}^\pm$. Let us first note that on $\mathscr{I}^\pm$, i.e. for $\Omega = 0$, eq. (10.55) implies

$$\hat{g}_{\mu\nu}\hat{\Delta}\,\Omega = 4\hat{\Delta}_\mu \hat{N}_\nu, \tag{10.58}$$

so that (10.53), or $\hat{\nabla}_\mu \hat{N}^\mu = 0$ on $\mathscr{I}^\pm$, is equivalent to the seemingly stronger condition

$$\hat{\nabla}_\mu \hat{N}_\nu = 0, \tag{10.59}$$

still on $\mathscr{I}^\pm$ only. This condition, in turn, implies that on $\mathscr{I}^\pm$ we have the geodesic equation

$$\hat{\nabla}_{\hat{N}} \hat{N} = 0. \tag{10.60}$$

In other words, in the "gauge" (10.53) the flow of the vector field $\hat{N}$, restricted to $\mathscr{I}^\pm$, consists of lightlike geodesics, and clause 3 requires that *these* particular lightlike geodesics be complete.[501]

Towards showing that (10.53) can be satisfied by a suitable choice of the the conformal factor Ω, we first relabel $\hat{g}$ as $\tilde{g}$, with ensuing differential operators $\tilde{\nabla}$ and $\tilde{\Delta}$, also relabel the original Ω as $\tilde{\Omega}$, i.e. $\tilde{g} = \tilde{\Omega}^2 g$, with $\tilde{N} = \tilde{\nabla}\tilde{\Omega}$, and define $\Omega = \omega\tilde{\Omega}$, where $\omega : \hat{M} \to (0,\infty)$ is smooth and nonzero on $\mathscr{I}^\pm$, for otherwise (10.61) below would make $\hat{N} = 0$, against Definition 10.1. Still using the notation (10.57), a straightforward computation shows that *on* $\mathscr{I}^\pm$ we have

$$\hat{N}_\mu = \omega \tilde{N}_\mu; \tag{10.61}$$

$$\hat{\nabla}_\mu \hat{N}_\nu = \omega \tilde{\nabla}_\mu \tilde{N}_\nu + \tilde{g}_{\mu\nu}\tilde{N}^\rho \tilde{\nabla}_\rho \omega. \tag{10.62}$$

Eq. (10.61) follows from $\hat{N}_\mu = \partial_\mu \Omega = \partial_\mu(\omega\tilde{\Omega}) = (\partial_\mu \omega)\tilde{\Omega} + \omega\partial_\mu\tilde{\Omega}$, which on $\mathscr{I}^\pm$, where $\tilde{\Omega} = 0$, equals $\omega\partial_\mu\tilde{\Omega} = \omega\tilde{N}_\mu$. Eq. (10.62) follows from once (covariantly) differentiating (10.55) and (10.56).[502] On $\mathscr{I}^\pm$, eqs. (10.62) and (10.58), but now for the "tilde" quantities, give

$$\hat{\nabla}_\mu \hat{N}_\nu = \tfrac{1}{4}\tilde{g}_{\mu\nu}(\omega\tilde{\Delta}\tilde{\Omega} + 4\tilde{N}^\rho \partial_\rho \omega). \tag{10.63}$$

Since $\tilde{N}^\rho \partial_\rho$ differentiates along $\mathscr{I}^\pm$, one can solve the ODE

$$\tilde{N}^\rho \partial_\rho \omega = -\tfrac{1}{4}\tilde{\Delta}\tilde{\Omega} \tag{10.64}$$

on $\mathscr{I}^\pm$ for given $\tilde{\Omega}$, with any initial condition ω_0, and this choice of ω achieves (10.59) and hence (10.53). Because of Definition 10.2.1, the initial condition may be stated on some fiducial copy of S^2, call it S^2_0. Furthermore, as a result of the classification of compact Riemann surfaces,

[501] Completeness of curves depends on their parametrization. Geodesics are affinely parametrized by definition (and an affine reparametrization does not affect their (in)completeness), but a change in Ω changes the metric and hence the notion of a geodesic with resect to $\hat{g}$ (for given g), so that completeness does depend on the choice of Ω.

[502] See e.g. Wald (1984), §11.1, Stewart (1991), §3.6, or Reall (2020), §5.2. for details. The extra derivative in the derivation of (10.62) requires better asymptotics than in Definition 10.2.2, such as $R_{\mu\nu} = O(\Omega^4)$ or $o(\Omega^5)$, se we assume this here. The following analysis is taken from Wald (1984), pp. 279–280, see also Reall (2020), §5.2.

in this case of genus zero, any Riemannian metric on S^2 is conformal to the standard one g_{S^2}. We may therefore choose ω_0 such that $g_{S^2_0} = g_{S^2}$. We now show that on the identification

$$\mathscr{I}^{\pm} \cong \mathbb{R} \times S^2 \tag{10.65}$$

from Definition 10.2.1, this remains true for all copies of S^2 in $\mathscr{I}^{\pm}$. Below we take $\mathscr{I}^{+}$; the other case $\mathscr{I}^{-}$ involves some sign changes. We first choose coordinates (u,θ,φ) on $\mathscr{I}^{+}$ such that the point $\gamma(u)$ labels the solution γ of (10.60) starting at $(\theta,\varphi) \in S^2_0$ for $s = 0$ with $\dot{\gamma}(0) = \hat{N}$ (so that $u \in \mathbb{R}$ by Definition 10.2.3). Because of (10.12), we can also use Ω as a coordinate on $\hat{M}$, at least near $\mathscr{I}^{+}$, so that we have local coordinates $(\Omega,u,\theta,\varphi)$. Note that

$$\frac{\partial}{\partial s} = \hat{N} = \hat{\nabla}\Omega, \tag{10.66}$$

which is a lightlike vector, is *tangent* to $\mathscr{I}^{+}$, whereas the vector field $\partial/\partial\Omega$ points *away from it*.

Eq. (10.59), written in terms of the Christoffel symbols, then implies that on $\mathscr{I}^{\pm}$, i.e. for $\Omega = 0$, the (θ,φ) components of $\hat{g}_{\mu\nu}$ are independent of u. Collecting all we know, we obtain

$$\hat{g}_{\Omega=0} = 2du\,d\Omega + g_{S^2}. \tag{10.67}$$

If we then introduce a–by definition–radial coordinate $v := 2/\Omega$, the physical metric near $\mathscr{I}^{\pm}$ is

$$g = -2du\,dv + \tfrac{1}{4}(v-u)^2 g_{S^2} + \cdots \tag{10.68}$$

as $v \to \infty$ at fixed u, where compared with (10.67) we have written $(v-u)^2$ instead of v^2. This is because, using (10.59), as $v \to \infty$ the remainder terms denotes by $\cdots$ can be shown to be:

$$\begin{aligned} &O(v) \text{ in } d\theta^2, d\varphi^2, d\theta\,d\varphi; && O(1) \text{ in } du^2, du\,d\theta, du\,d\varphi; \\ &O(1/v) \text{ in } dv\,du, dv\,d\theta, dv\,d\varphi; && O(1/v^3) \text{ in } dv^2. \end{aligned} \tag{10.69}$$

Hence the leading terms of g near $\mathscr{I}^{+}$ are the same as in Minkowski space, cf. (10.16).

This completes the exegesis of Definition 10.2. Having used Minkowski space-time to motivate this definition, let us use the Schwarzschild solution to check that it is reasonable. To find a conformal completion of the Schwarzschild space-time (9.46) with metric (9.45) in *outgoing* Eddington–Finkelstein coordinates (u,r,θ,φ), first change r to $w = 1/r$, which gives

$$g = \frac{1}{w^2}\left(2du\,dw - w^2(1-2mw)dw^2 + d\theta^2 + \sin^2\theta\, d\varphi^2\right). \tag{10.70}$$

Then take $\Omega(u,w,\theta,\varphi) = w$, which obviously gives the unphysical metric

$$\hat{g} = 2du\,dw - w^2(1-2mw)dw^2 + d\theta^2 + \sin^2\theta\, d\varphi^2, \tag{10.71}$$

defined on a manifold with boundary $\hat{M}_S$ given by adding all points with $w = 0$. The map $i : M_S \hookrightarrow \hat{M}_S$ is then the identity, much as in the conformal compactification of the Poincaré disc $\mathbb{D}$ reviewed in §10.1. Since $r \to \infty$ at fixed u amounts to $v \to \infty$, this procedure only adds future null infinity $\mathscr{I}^{+}$. To add past null infinity $\mathscr{I}^{-}$, one should repeat this procedure for the *incoming* Eddington–Finkelstein coordinates (v,r,θ,φ), and as long as $r > 2m$ these can be combined to define a conformal completion of the corresponding part of M_S, where the passage from outgoing to incoming coordinates is just a coordinate transformation. However, we will not spell out the result, since what we really are interested in is the entire region $0 < r < \infty$, where the two coordinate systems are no longer related by a coordinate transformation.

To define a conformal completion of all of M_S, we may enlarge it to the Kruskal space-time M_K, described in (U,V) coordinates. In a subtle variation on (10.21) - (10.22), we define

$$V = \sinh(\tan P); \qquad P = \arctan(\operatorname{arcsinh} V); \tag{10.72}$$

$$U = \sinh(\tan Q); \qquad Q = \arctan(\operatorname{arcsinh} U); \tag{10.73}$$

$$P,Q \in (-\tfrac{1}{2}\pi, \tfrac{1}{2}\pi); \qquad P+Q \in (-\tfrac{1}{2}\pi, \tfrac{1}{2}\pi). \tag{10.74}$$

where the last condition is necessary to keep $r > 0$. The conformal completion then has $P,Q \in [-\frac{1}{2}\pi, \frac{1}{2}\pi]$, still subject to the second part of (10.74); for the upper $r = 0$ branch in the Kruskal diagram (which is part of neither M_K not $\hat{M}_K$) corresponds to $P+Q = \frac{1}{2}\pi$ in the Penrose diagram, whereas the lower $r = 0$ branch is $P+Q = -\frac{1}{2}\pi$. This gives

$$g = -\frac{32m^3 e^{-r/2m}}{r} \cdot \frac{\cosh(\tan P)\cosh(\tan Q)}{\cos^2 P \cos^2 Q} dP\,dQ + r^2(d\theta^2 + \sin^2\theta d\varphi^2), \tag{10.75}$$

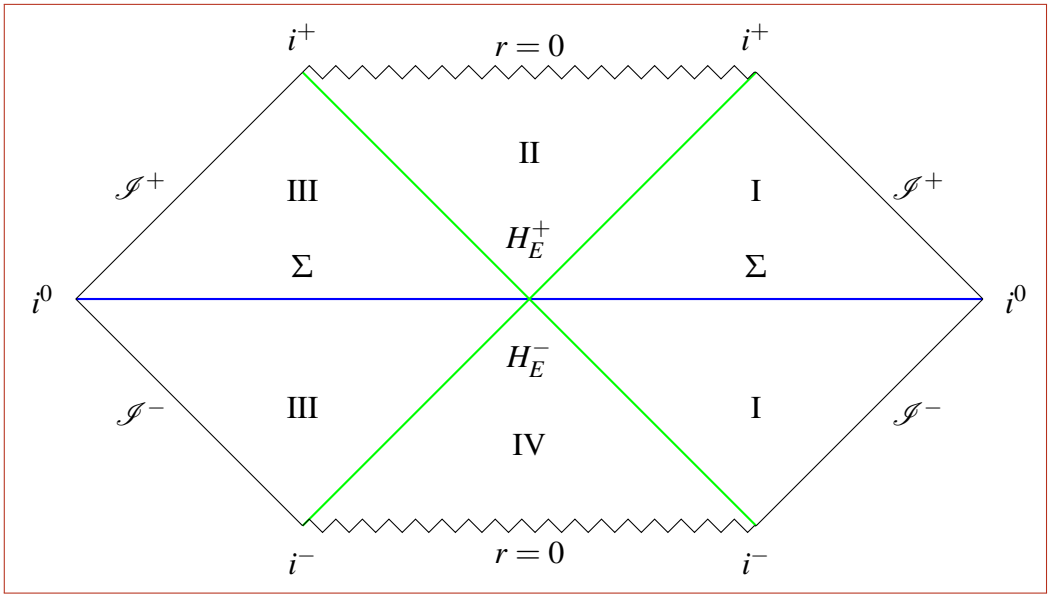
where r is regarded as a function of P and Q similar to the explanation following (9.55), now adding (10.72) - (10.73) to the story. For the conformal factor we now obviously take

$$\Omega(P,Q)^2 = \frac{re^{r/2m}}{32m^3} \frac{\cos^2 P \cos^2 Q}{\cosh(\tan P)\cosh(\tan Q)}, \tag{10.76}$$

which has the right asymptotics $\Omega(r) \sim 1/r$ as $r \to \infty$.[503] The rescaled metric then becomes

$$\hat{g} = -dP\,dQ + r^2\Omega(P,Q)^2(d\theta^2 + \sin^2\theta d\varphi^2), \tag{10.77}$$

which shows that the two-spheres S^2 in $\mathscr{I}^\pm$ acquire their usual metric g_{S^2}. A simple computation shows that Definition 10.2.3 is satisfied. The other two clauses are obvious from (9.46), which also applies to M_K, and from the fact that the Ricci tensor of g vanishes.



Penrose diagram for Kruskal space-time M_K, in which Schwarzschild space-time M_S corresponds to regions I and II (excluding the SE–NW green diagonal but including the upper half of the green SW–NE line), see §9.4. The P-axis is at 45° and the Q-axis is at 135°, just like V and U in the Kruskal diagram, or (p,q) in the Minkowski case. Hence radial lightlike geodesics move parallel to these axes. The green lines are event horizons, whilst the blue line is a Cauchy surface.

[503] To see this, use (9.56), which for $r \to \infty$ gives $\Omega^2 \sim \tanh(\tan P)\tanh(\tan Q)\cos^2 P \cos^2 Q$. Towards e.g. $\mathscr{I}^+$, where $P \to \frac{1}{2}\pi$ at fixed Q, this gives $\Omega \sim \cos P$. In the same regime, $r \sim \ln V \sim \ln(\sinh(\tan P)) \sim \tan P \sim 1/\cos P$.

- **Future null infinity** $\mathscr{I}^+$ *has two components: on the right it has* $P = \frac{1}{2}\pi$, $Q \in (-\frac{1}{2}\pi, 0)$, *times the two-sphere* S^2. *For the Schwarzschild space-time* M_S *this is all. For Kruskal space-time,* $\mathscr{I}^+$ *in addition has the component on the left, where* $P \in (-\frac{1}{2}\pi, 0)$, $Q = \frac{1}{2}\pi$.
- **Past null infinity** $\mathscr{I}^-$ *similarly has two components: on the right it has* $P \in (0, \frac{1}{2}\pi)$, $Q = -\frac{1}{2}\pi$ *(which is all for* M_S*), and on the left,* $P = -\frac{1}{2}\pi$, $Q \in (0, \frac{1}{2}\pi)$.
- **Future timelike infinity** i^+ *is* $(P = \frac{1}{2}\pi, Q = 0)$ *on the right and* $(0, \frac{1}{2}\pi)$ *on the left.*
- **Past timelike infinity** i^- *is* $(P = 0, Q = -\frac{1}{2}\pi)$ *on the right and* $(-\frac{1}{2}\pi, 0)$ *on the left.*
- **Spacelike infinity** i^0 *is* $(\frac{1}{2}\pi, -\frac{1}{2}\pi)$ *on the right and* $(-\frac{1}{2}\pi, \frac{1}{2}\pi)$ *on the left.*[504]

We return to the general theory. The following definition is due to Hawking and Penrose.[505]

Definition 10.4 *Let* (M, g) *be a space-time that is asymptotically flat at null infinity. The* **black hole region** B^+ *and the* **white hole region** B^- *in* M *are defined (and then re-expressed) by*

$$B^+ := M \backslash J^-(\mathscr{I}^+) = M \backslash I^-(\mathscr{I}^+); \qquad B^- := M \backslash J^+(\mathscr{I}^-) = M \backslash I^+(\mathscr{I}^-). \tag{10.78}$$

Each connected component of $B^\pm$*, if not empty, is a* **black hole** *(or* **white hole***). The boundaries*

$$H_E^+ := \partial B^+ = \partial I^-(\mathscr{I}^+); \qquad H_E^- := \partial B^- = \partial I^+(\mathscr{I}^-) \tag{10.79}$$

decompose into the **future** *and* **past event horizons** *of each of the black and white holes in* M.

The hole regions $B^\pm$ are closed, so that $H_E^\pm \subset B^\pm$, i.e. *the event horizons form part of the holes.*

For Minkowski space-time $(\mathbb{M}, \eta)$with conformal completion $(\hat{\mathbb{M}}, \hat{\eta})$, as in (10.24) - (10.25),

$$J^+(\mathscr{I}^-) \cap \mathbb{M} = J^-(\mathscr{I}^+) \cap \mathbb{M} = \mathbb{M}; \qquad B^\pm = \emptyset. \tag{10.80}$$

Thus Minkowski space-time is free of holes.[506] For Kruskal space-time (M_K, g_K), the Penrose diagram shows that $J^-(\mathscr{I}^+) \cap M_K$ consists of regions I, III, and IV (without boundaries), whilst $J^-(\mathscr{I}^+) \cap M_S$ is just region I. In both cases the black hole B^+ is in region II (with boundaries).

---

[504] Eqs. (10.72) - (10.73) are suggested by Penrose (1968), p. 209. The choice $V = \tan P$, $U = \tan Q$, as in Valiente Kroon (2016), p. 165, will not work here; his metric $\mathbf{g}_{\mathscr{I}}$ (which in our notation would be $\hat{g}$) vanishes on $\mathscr{I}^\pm$.

[505] See Penrose (1969) and Hawking (1972), as further analyzed in footnote 485. The non-defining equalities in (10.78) and (10.79) follow from $J^\mp(\mathscr{I}^\pm) \cap M = I^\mp(\mathscr{I}^\pm) \cap M = I^\mp(\mathscr{I}^\pm)$, where the last equality follows because $\mathscr{I}^\pm$ is null, cf. Proposition 10.3 and Lemma 4.16. To prove the first equality, we take $x \in J^-(y)$ for some $x \in M$ and $y \in \mathscr{I}^+$. Then we are ready if $x \in I^-(y)$, so assume $x \in J^-(y) \backslash I^-(y)$, in which case x and y must be connected by a lightlike (pre)geodesic, cf. Corollary 5.14.1. By Propositions 10.3 and 6.9 one may take any point $z \in \mathscr{I}^+$ on a fd lightlike geodesic through y within $\mathscr{I}^+$, which can clearly be connected to x by a path that is not a lightlike (pre) geodesics. Hence $x \in I^-(z)$ again by Corollary 5.14.1. See also Wald (1984), p. 308. Finally, if $Y \subset X$, then ∂Y consists of all $x \in X$ for which any nbhd U intersects both Y and $X \backslash Y$. Hence $\partial Y = \partial(X \backslash Y)$.

[506] On the other hand, truncating $\mathscr{I}^+$ to for example $\{(p, q, \theta, \varphi) \mid p = \pi/2, q \in (-\pi/2, 0)\}$ instead of (10.26), would turn $J^-(\mathscr{I}^+) \cap \mathbb{M}$ into the region $u < 0$ and hence make the future lightcone $J^+(0)$ a fake black hole in $\mathbb{M}$. This would still satisfy Definition 10.1, but Definition 10.2.3 would now fail. Removing $B^+ = J^+(0)$, the space-time $(M, g) := (\mathbb{M} \backslash J^+(0), \eta)$ still has a conformal completion (for example the one just described), and is now free of black holes, but its future null infinity is incomplete. Excluding cases like this was in fact what led Geroch & Horowitz (1978) to introduce clause 3 in Definition 10.2 (in slightly different form). The inextendibility condition proposed by Geroch (1977) fails to exclude cases like $(\mathbb{M} \backslash J^+(0), \eta)$, but his inextendibility plus some regularity condition did enable him to prove uniqueness of conformal completions, a result that seems to have no analogue for Definition 10.2. See also Chruściel (2002), §3.4 for further comments on this issue.

The future event horizon H_E^+ consists of the two upper $r = 2m$ lines. Similarly, $J^+(\mathscr{I}^-) \cap M_K$ consists of regions I, II, and III (without boundaries), and $J^+(\mathscr{I}^-) \cap M_S$ is regions I and II, i.e., all of M_S. The white hole B^- in M_K is region IV (with boundaries), with past event horizon H_E^- consisting of the two lower green $r = 2m$ lines. See also the Penrose diagram below (10.77).

To close this section, we discuss the fact that though mathematically sweet, Definition (10.79) of a horizon, based on the idealizations in Definition 10.2, is not entirely uncontroversial:

> This definition depends on the whole future behaviour of the solution (...) one cannot find where the event horizon is without solving the Cauchy problem for the whole future development of the [partial Cauchy] surface.' (Hawking & Ellis, 1973, p. 319)
>
> This definition is global in a strong and straightforward sense: the idea that nothing can escape the interior of a black hole once it enters makes implicit reference to all future time–the thing can never escape no matter how long it tries.[507] Thus, in order to know the location of the event horizon in spacetime, one must know the entire structure of the spacetime, from start to finish, so to speak, and all the way out to infinity. As a consequence, no local measurements one can make can ever determine the location of an event horizon. (...) Another disturbing property of the event horizon, arising from its global nature, is that it is prescient. Where I locate the horizon today depends on what I throw in it tomorrow–which future-directed possible paths of particles and light rays can escape to infinity starting today depends on where the horizon will be tomorrow, and so that information must already be accounted for today. Physicists find this feature even more troubling. (Curiel, 2019b, p. 29)
>
> [The notion of a horizon] is probably very useless, because it assumes we can compute the future of real black holes, and we cannot. (Carlo Rovelli, quoted in Curiel, 2019b, p. 30)
>
> I have no idea why there should be any controversy of any kind about the definition of a black hole. There is a precise, clear definition in the context of asymptotically flat spacetimes (...) I don't see this as any different than what occurs everywhere else in physics, where one can give precise definitions for idealized cases but these are not achievable/measurable in the real world. (Bob Wald, quoted in Curiel, 2019b, p. 32)

However, the disagreement may not be so bad, since two kinds of idealizations are involved here: (i) The ability to know an entire space-time (M, g), either from initial data or by direct construction, and (ii) The construction of (null) *infinity* from which the black hole and its event horizon are defined. Rovelli's comment seems to apply to the first point and Wald's to the second. On the other hand, the second point is predicated on the first, which remains unresolved, except from the point of view of a Laplacian demon. Definition (10.79) is an axiomatic approach to black holes, subject to Russell's famous charge that 'The method of "postulating" what we want has many advantages; they are the same as the advantages of theft over honest toil.' However, nothing is wrong with an axiomatic approach as long as one can find representative and realistic models for the axioms (or definitions) that show that they are reasonable. This is certainly the case here, where the known exact black hole solutions validate all definitions.

In any case, discussions like this have led to alternative, more local definitions of event horizons, of which *apparent*, *dynamical*, *isolated*, and *naive* horizons are examples.[508] For stationary black holes one may even avoid $\mathscr{I}$ at no cost in defining event horizons, see §10.9.

[507] Or, as Ashtekar & Galloway (2005), p. 2, insightfully write in an article on dynamical horizons: '$[H_E^+]$ is the boundary of an interior spacetime region from which causal signals can never be sent to the asymptotic observers, *no matter how long they are prepared to wait.* The region is therefore "black" in an absolute sense.'

[508] See e.g. Hawking & Ellis (1973), §9.2, Ashtekar & Krishnan (2004), Booth (2005), Chruściel (2002; 2020, §8.4), Hayward (2013), and Faraoni (2015). See also §10.11 for a short introduction to apparent horizons.

10.4 Cosmic censorship à la Penrose

A key issue in the theory of black holes is Penrose's ***cosmic censorship*** conjecture, which he first raised in 1969 and which underwent several refinements, bifurcations, and reformulations since then.[509] One way to understand this development is to compare the actual achievement of Penrose's 1965 incompleteness theorem with its intended goal. Quoting Penrose's 2020 Physics Nobel Prize citation, this goal was to prove that 'black hole formation is a robust prediction of the general theory of relativity' (see also chapter 6). However, what the theorem proved was that the conjunction of (i) a non-compact Cauchy surface; (ii) the null energy condition, and (iii) the presence of a trapped surface, implies lightlike geodesics incompleteness (cf. Theorem 6.15). In the light of the analysis in the preamble to chapter 6, it is clear that two things were missing:

1. To get closer to a curvature singularity as the source of lightlike geodesic incompleteness, one should get rid of the possibility of extendibility of the space-time in question.

2. Although an event horizon is what makes black holes black, this concept plays no role whatsoever in the theorem, and hence it should be involved one way or the other.[510]

Briefly, the first point leads to ***strong cosmic censorship***, whereas the second leads to ***weak cosmic censorship***.[511] The latter came first. Here is Penrose's original formulation:

> We are thus presented with what is perhaps the most fundamental question of general-relativistic collapse theory, namely: does there exist a "cosmic censor" who forbids the appearance of naked singularities, clothing each one in an absolute event horizon? In one sense, a "cosmic censor" can be shown *not* to exist. For it follows from a theorem of Hawking that the "big bang" singularity is, in principle, observable. But it is not known whether singularities observable from outside will ever arise in a generic *collapse* which starts off from a perfectly reasonable nonsingular initial state. (Penrose, 1969, p. 1162)

Or, in Penrose (1979), p. 618, with an emphasis on nitial data:

> A system which evolves, according to classical general relativity with reasonable equations of state, from generic non-singular initial data on a suitable Cauchy hypersurface, does not develop any spacetime singularity which is visible from infinity. (Penrose, 1979, p. 618).

Visibility from infinity, then, is blocked by an event horizon. However, Penrose argued:

> It seems to me to be comparatively unimportant whether the observer himself can escape to infinity. Classical general relativity is a scale-invariant theory, so if locally naked singularities occur on a very tiny scale, they should also, in principle, occur on a very large scale in which a 'trapped' observer could have days or even years to ponder upon the implications of the uncertainties introduced by the observations of such a singularity. (...) It would seem, therefore, that if cosmic censorship is a principle of Nature, it should be formulated in such a way as to preclude such *locally* naked singularities. (Penrose, 1979, p. 619)

[509] See Earman (1995), chapter 2, for a complete survey up to that point, and Joshi (1993, 2007) for case studies.

[510] Yet in one of the most important papers about black holes in observational astronomy (Event Horizon Telescope Collaboration, 2019a), which would have deserved to share the 2020 Physics Nobel Prize with Penrose, the authors write: 'A defining feature of black holes is their event horizon, a one-way causal boundary in spacetime from which not even light can escape (Schwarzschild 1916). The production of black holes is generic in GR (Penrose 1965).'

[511] In 1965 Penrose knew the concept of future null infinity $\mathscr{I}^+$, since he had conceived it himself at least a year earlier (Penrose, 1964). We now know that this leads to a clean definition of black holes and their (absolute) event horizons (see §10.3), but at the time $\mathscr{I}^+$ was apparently supposed to be relevant mainly for the study of gravitational radiation. Its application to black holes had to wait until at least Penrose (1969). See footnote 485.

This preclusion, then, is Penrose's (original) idea of ***strong cosmic censorship***. Although historically the strong version postdated the weak one, it is conceptually easier to start defining strong cosmic censorship rigorously, and then move to the former as a modification thereof.[512]

The problem is to define what it means for a signal to emanate from a singularity, since the latter is not part of space-time. To resolve this, we recall that in the context of the singularity/incompleteness theorems, singularities were tentatively captured by incomplete causal geodesics in space-time, and hence one would expect Penrose to use these here, too. However, instead of *incomplete causal geodesics* he now uses *inextendible causal curves*.[513] It turns out that the change from 'causal' to 'timelike' does not matter,[514] but the change from *geodesics* to *curves*, which is required by the proof of Penrose's Theorem 10.6 below,[515] is substantial.[516]

Now the reasoning that leads to a definition of strong cosmic censorship is as follows.

1. If (M,g) is strongly causal,[517] then $I^-(x) = I^-(x')$ iff $x = x'$. This allows us to exchange properties of points x for properties of their timelike pasts $I^-(x)$, which will be crucial.

2. By definition, we have $z \ll x$, i.e. $z \in I^-(x)$, iff some future-directed timelike curve from z can reach x. In this case z can signal to x, or influence x, and since x can "see" z, we say that z *is locally naked for* x. This, then, is equivalent to the property
$$I^-(z) \subset I^-(x). \tag{10.81}$$

3. If z is the endpoint of some f-d timelike curve c, then $I^-(z) \subset I^-(x)$ iff $I^-(c) \subset I^-(x)$.

4. The point is that this also works if c has no endpoint and hence defines a "singularity": this singularity is deemed locally naked for x iff $I^-(c) \subset I^-(x)$. In summary:

[512] The following discussion is based on Penrose (1979), which we simplify by removing the TIPs of Geroch, Kronheimer, & Penrose (1972) from the discussion, including the proof of Theorem 10.6. Penrose's timelike curves are smooth (Penrose, 1972, pp. 2–3), whereas we work with *continuous causal* curves, see Definition 5.20 in §5.6.

[513] (In)completeness of non-geodesic curves depends on the parametrization. If continuous causal curves are parametrized by arc length as measured by an auxiliary complete Riemannian metric (see footnote 512), then any inextendible curve has infinite arc length, see Lemma 5.22. Also, recall that (affinely parametrized) timelike *geodesics* are incomplete iff they are inextendible and have finite parameter length, cf. Proposition 5.19.

[514] In the light of the analysis below, this follows from Theorem (2.3) in Geroch, Kronheimer, & Penrose (1972). Causal *geodesics* would presumably lead to some weaker causality condition than global hyperbolicity.

[515] It is the second ('converse') part of the proof of Theorem 10.6 below that does not work for causal geodesics instead of curves, since the curve c constructed there is not necessarily a geodesic. This goes back to the definition of domains of dependence and Cauchy surfaces in terms of causal curves rather than geodesics.

[516] To bridge the gap with incomplete geodesics, note that, heuristically, an inextendible causal curve may either go off to infinity, or hover around in a compact set, or stop at some boundary of an extendible space-time, or hit something like a curvature singularity. The first possibility is not excluded *a priori*, but seems hard to combine with the key condition (10.82) below, and according to Penrose, in space-times that are asymptotically flat at null infinity this is even impossible. See Penrose (1979), p. 623. For example, condition (10.82) below may be satisfied in anti-de Sitter space, which is hardly singular. But anti-de Sitter space has a negative cosmological constant with timelike future null infinity. Secondly, hovering around in a compact set is impossible in a strongly causal space-time. If we therefore assume that our space-time is both strongly causal and asymptotically flat at null infinity, as is typically the case for black hole space-times, then we are left with inextendible causal curves that may either lead to the edge of an extendible space-time or crash into a singularity. Therefore, under the stated assumptions the situation only differs from the one in the singularity theorems in that our curves are not necessarily geodesics. Asymptotic flatness is not assumed in either Definition 10.5 or Theorem 10.6, so that global hyperbolicity excludes the local nakedness of even more singularities than those described by the incompleteness theorems.

[517] The following property makes a space-time ***past distinguishing***, which on the causal ladder is well below strong causality (Minguzzi, 2019, chapter 4, Definition 4.46). However, through its implication of non-total imprisonment, strong causality is also used through invocation of Theorem 2.53 in Minguzzi (2019) in our proof of Theorem 10.6.

Definition 10.5 *A strongly causal space-time (M,g) contains a* **locally naked singularity** *if there is a future-directed future-inextendible causal curve c in M, and a point $x \in M$, such that*

$$I^-(c) \subset I^-(x). \tag{10.82}$$

Penrose's **strong cosmic censorship conjecture** *states that "generic" [in his own words: "physcially reasonable"] space-times do not contain locally naked singularities.*[518]

The curve c defines or represents this "singularity", and $x \in M$ is in its chronological future.[519] For example, in Minkowski space-time, take $z \in I^-(x)$ and remove z. Then any fd future-inextendible timelike curve c whose endpoint would have been z satisfies (10.82). Of course, it should be defined precisely what "generic" means, lest these conjectures turn into a definition of genericity! Penrose did not do this, but we will return to this point in §10.5. The following theorem, due to Penrose (1979), characterizes–or redefines–his idea of strong cosmic censorship.

Theorem 10.6 *A strongly causal space-time has no locally naked singularities, i.e. satisfies strong cosmic censorship, iff it is globally hyperbolic.*

Proof.[520] We prove the inference from global hyperbolicity to the absence of locally naked singularities by contradiction.[521] Suppose that (M,g) is globally hyperbolic and that (10.82) holds for some c and x. Take y on c and then a future-directed sequence (y_n) of points on c, with $y_0 = y$. Because of (10.82) this sequence lies in $J^+(y) \cap J^-(x)$, which is compact by assumption. Hence (y_n) has a limit point z in $J^+(y) \cap J^-(x)$. Now define curves (γ_n) as the segments of c from y to y_n. By Lemma 5.26, these curves have a uniform limit γ. Its arc length (as measured by an auxiliary complete Riemannian metric, see footnote 512) is, on the one hand, infinite (since c is endless and hence has infinite arc length, which is approached as the y_n move up along c). But on the other hand it is finite, since γ must end at z (and fd continuous causal curves have finite arc length iff they have an endpoint). Hence (10.82) cannot be true and the inference is ready.

The (contrapositive) proof of the converse implication relies on the following lemma:[522]

Lemma 10.7 *Let (M,g) be a space-time, let $S \subset M$ be closed and achronal, and let $x,y \in M$.*

1. *If $y \in \mathrm{int}(D^-(S))$, then $J^+(y) \cap J^-(S)$ is compact. In particular, taking $S = \partial I^-(x)$ and assuming $y \in I^-(x)$, it follows that $J^+(y) \cap J^-(x)$ is compact.*
2. *We have $\mathrm{int}(D^-(S)) = I^-(S) \cap I^+(D^-(S))$.*

[518] One may also use *past-directed* inextendible causal curves, replacing $I^-(\cdot)$ by $I^+(\cdot)$, etc. For strong cosmic censorship this definition is equivalent to the given one, as follows from Theorem 10.6 below.

[519] One may also regard c or rather $I^-(c)$ as an ***ideal point*** of space-time. Assuming strong causality, Geroch, Kronheimer, & Penrose (1972) and in their wake Hawking & Ellis (1973), §6.8, show that both real points and ideal points of M correspond to subsets $U \subset M$ that are: (i) open, (ii), past sets, i.e. $I^-(U) \subset U$, and (iii) indecomposable, in that $U \neq U_1 \cup U_2$ where U_1 and U_2 have properties (i) and (ii) and are neither empty nor equal to U. Such sets are called IP (for Indecomposable Past set), and those that are not of the form $U = I^-(x)$ for some $x \in M$ are TIPs (for Terminal IPs); these TIPs are $U = I^-(c)$ for some future-inextendible timelike curve c.

[520] A heuristic argument for part 2 of the theorem is that a locally naked singularity, represented by c as in Definition 10.5, will not reach any wannabe Cauchy-surface Σ in $I^+(x)$, since it crashes at the singularity lying in the past of x. Conversely, if no Cauchy surface exists then one can construct such a curve c. See also §10.5 below.

[521] Penrose (1979) gives a considerably more complicated argument, in terms of his TIP's (which we avoid).

[522] We follow Penrose (1979), p. 624. The lemma combines Propositions 5.20 and 5.5 (h) in Penrose (1972).

The first point is a variation on Proposition 5.39, in which $D(S)$ is replaced by $D^-(S)$. The specification follows from the property $J^-(\partial I^-(x)) = J^-(x)$. The second point is a simple consequence of the definitions in question. To prove the converse direction of Theorem 10.6 (contrapositively), assume that (M,g) is not globally hyperbolic. Then, under the assumption of strong causality, there are x,y for which $J^-(x)\cap J^+(y)$ is not compact (cf. Definition 5.27, where strong causality implies non-imprisoning). In view of (5.96) we may assume that $y\in I^-(x)$. Part 1 of Lemma 10.7 gives $y\notin \mathrm{int}(D^-(\partial I^-(x)))$. Part 2 gives some $y'\in I^-(x)$ with $y'\notin D^-(\partial I^-(x))$, so that, by definition of D^-, there exists some fd future-inextendible curve c from y' that avoids $\partial I^-(x)$. Since $y'\in I^-(x)$, this curves does lie in $I^-(x)$, and hence (10.82) holds.[523] □

We now (ahistorically) move from strong to weak cosmic censorship. If our space-time (M,g), so far merely assumed strongly causal, is also asymptotically flat at null infinity, then Definition 10.5 can be modified by requiring $x\in I^-(\mathscr{I}^+) = I^-(\mathscr{I}^+)\cap M = J^-(\mathscr{I}^+)\cap M$. This means that the "singularity" represented by the inextendible curve c is hidden from observers in $I^-(\mathscr{I}^+)$, although it may be "naked" to observers in the black hole region B^+, cf. (10.78).

Definition 10.8 *1. A strongly causal space-time (M,g) that is asymptotically flat at null infinity contains a* **naked singularity** *if there is a future-directed future-inextendible causal curve c in M and a point $x\in I^-(\mathscr{I}^+)$ such that $I^-(c)\subset I^-(x)$.*

2. The **weak cosmic censorship conjecture** *states that no strongly causal space-time arising from "generic" smooth and complete initial data contains a naked singularity.*

A slight change in the proof of Theorem 10.6, involving a case distinction on c, yields:

Theorem 10.9 *Weak cosmic censorship holds iff $I^-(\mathscr{I}^+)$ is globally hyperbolic.*

In our formulation, strong cosmic censorship implies weak cosmic censorship.[524] A slight variation of Definition 10.8, which is relevant for black hole thermodynamics and uniqueness theorems (see §10.10 - §10.12), has the same virtue. The ***domain of outer communication*** $\mathrm{DOC}(\mathscr{I})$ of a space-time that is asymptotically flat at null infinity is defined by

$$\mathrm{DOC}(\mathscr{I}) := I^-(\mathscr{I}^+)\cap I^+(\mathscr{I}^-). \tag{10.83}$$

If we now replace the condition $x\in I^-(\mathscr{I}^+)$ in Definition 10.8 by $x\in \mathrm{DOC}(\mathscr{I})$, we obtain a slightly different formulation of weak cosmic censorship; let us call it DOC–WCC. On this definition, DOC–WCC at least *implies* that $\mathrm{DOC}(\mathscr{I})$ is globally hyperbolic (there is no "iff").

Of course, by changing '$x\in I^-(\mathscr{I}^+)$' in Definition 10.8 to '$x\in R$' for some causally interesting region $R\subset M$, or even $R\subset \hat{M}$, one can engineer the definition of weak cosmic censorship in any desired way,[525] preferably with some corresponding version of Theorem 10.9.

[523] All this can be checked in the Minkowskian example following Definition 10.5, where, assuming $z\in I^+(y)$, the removal of z ruins compactness of $J^+(y)\cap J^-(x)$ and hence global hyperbolicity. The existence of c is trivial.

[524] This follows from the definitions, but it also follows from Theorems 10.6 and 10.9 and the observation that if (M,g) is globally hyperbolic, then so is $I^-(\mathscr{I}^+)$: if $x\in J^+(y)$ for $x,y\in I^-(\mathscr{I}^+)$, then $J^+(y)\cap J^-(x)\subset I^-(\mathscr{I}^+)$.

[525] For example, Tipler, Clarke & Ellis (1980), p. 176, define weak cosmic censorship as global hyperbolicity of $J^-(\mathscr{I}^+)\subset \hat{M}$, which does not follow from global hyperbolicity of $I^-(\mathscr{I}^+)\subset M$. See e.g. Chruściel, & Galloway (2019). Penrose's (1979) prose suggests replacing $I^-(\mathscr{I}^+)$ by $J^-(\mathscr{I}^+)\cap J^+(\Sigma)$, where Σ is some wannabe Cauchy surface in M. Hawking & Ellis, 1973, p. 312, say that (M,g) is ***future asymptotically predictable*** from Σ if the conformal completion $\hat{M}$ of M contains an open subset $\tilde{V}\subset\tilde{M}$ such that $\overline{J^-(\mathscr{I}^+)\cap M}\subset\tilde{V}$ and $(\tilde{V},\tilde{g})$ is globally hyperbolic. This is equivalent to the definition we just attributed to Penrose only under further regularity assumptions (Królak, 1986, Lemma 2.10). See also Wald (1984), §12.1 and Chruściel (2020), §3.5.1.

10.5 Cosmic censorship in the initial value (PDE) formulation

Let us pause to take stock, especially with regard to the two points laid out at the beginning of §10.4 that were claimed to be missing from Penrose's incompleteness/singularity theorem 6.15.

As to the second point, Theorem 6.15 has the following remarkable extension:

Corollary 10.10 *Under the assumptions of Penrose's Theorem 6.15, the "singularity" defined by the ensuing incomplete lightlike geodesic cannot be locally naked (let alone naked).*

This follows from Theorems 10.6 and 5.34, simply because Theorem 6.15 assumes a Cauchy surface. Hence it does not even seem necessary to postulate the existence of an event horizon that covers the singularity! However, the standard black hole space-times studied in chapter 9 do have event horizons, which, acting as one-way membranes, accomplish more than dressing the singularity. Thus Corollary 10.10 merely confirms the mismatch between Theorem 6.15 and the physical concept of a black hole, which is predicated on having an event horizon; the latter still needs to be postulated on top of the assumptions of Theorem 6.15. It therefore seems that weak cosmic censorship cannot be settled at the axiomatic level and has to be dealt with by means of (preferably "generic") case studies of black hole formation, so far with mixed results.[526]

For the first point in §10.4, at first sight strong cosmic censorship as in Definition 10.5 seems to have nothing to do with inextendibility. But in fact–and this must have been clear to Penrose all along–it has everything to do with it! But let us first cause further confusion by noting that in the light of Theorem 10.6, Penrose's Definition 10.5 of strong cosmic censorship is inappropriate if one adheres to the PDE approach to GR and especially to its second principle laid out in §7.6, namely that all valid questions in GR are questions about the MGHD (or maximal Cauchy development) $(M,g,\imath)$ of initial data $(\Sigma,\tilde{g},\tilde{k})$. Indeed, a MGHD is always globally hyperbolic, and hence strong cosmic censorship is automatic. However, this should disqualify neither Definition 10.5 nor the notion of a MGHD; it is their combination that seems a mismatch.

To overcome this–whilst admitting that there is no crystal-clear logical path from Penrose's formulationto the PDE version below–we introduce the following variation of Definition 7.4:[527]

Definition 10.11 *A* **development** *of initial data $(\Sigma,\tilde{g},\tilde{k})$–satisfying the vacuum constraints–is a triple $(M,g,\imath)$, where (M,g) is a space-time solving the vacuum Einstein equations, and $\imath:\Sigma\to M$ is an embedding such that $\imath^*g=\tilde{g}$ and $\imath(\Sigma)$ has extrinsic curvature $\tilde{k}$.*

[526] Christodoulou (1999b) proved his own–PDE–version of the weak cosmic censorship conjecture (Definition 10.13.2) for the spherically symmetric gravitational collapse of a scalar field, but on the basis of genericity conditions whose relevance has been questioned (Gundlach & Martin-Garcia, 2007, §3.4). See also the references in footnote 297. More generally, the status of weak cosmic censorship seems mixed also in earlier heuristic formulations in terms of an event horizon; see e.g. Joshi (1993, 2007), Królak (1999), and Ong (2020).

[527] This theory is due to Chruściel (1992). Earlier, Moncrief & Eardley (1981), p. 889, proposed an '(informally stated) global existence conjecture' stating that "Every asymptotically flat initial data set with $\mathrm{tr}K=0$ may be evolved to arbitrarily large times", adding that its proof would 'in essence prove the [weak] cosmic censorship conjecture for asymptotically flat space-times'. For initial data given on a compact Cauchy surface they propose something similar, and in doing so they opened the door to regarding cosmic censorship as a global existence problem for the Einstein equations. In this spirit, Moncrief (1981), p. 88, paraphrases Penrose's strong cosmic censorship as expressed by Theorem 10.6 as: 'the maximal Cauchy development of a generic initial data set is inextendible.' Similarly, Chruściel, Isenberg, & Moncrief (1990) open their abstract as follows: 'The strong cosmic censorship conjecture states that 'most' spacetimes developed as solutions of Einstein's equations from prescribed initial data cannot be extended outside of their maximal domains of dependence.' In §3 they further specify 'most' in terms of open and dense subsets in the space of initial data.

This may be adapted to the case with matter. The difference between a *development* and a *Cauchy development* is that in the former $\imath(\Sigma)$ is no longer required to be Cauchy surface in M, so that (M,g) is not necessarily globally hyperbolic. Such a development is called *maximal* if there is no extension (M',g') that also satisfies the vacuum Einstein equations, and one proves existence of maximal developments (but not uniqueness up to isometry, as in the globally hyperbolic case).

We now apply Penrose's strong cosmic censorship to such a maximal development, i.e. require it to be globally hyperbolic. The connection with inextendibility is then easily made:

Proposition 10.12 *The maximal development of given initial data is globally hyperbolic iff the* MGHD *of these data is is inextendible* as a solution to the vacuum Einstein equations.

Of course, "the" MGHD is only defined up to isometry, see Theorem 7.10, so be aware.

Proof. As explained in §7.6, the set of isometry classes $[M,g,i]$ of Cauchy developments (M,g,i) of given initial data $(\Sigma,\tilde{g},\tilde{k})$ is partially ordered, and by Theorem 7.12 the MGHD $[M_t,g_t,\imath_t]$ is its top element. Hence if some maximal development $(M_m,g_m,\imath_m)$ is globally hyperbolic then $[M_m,g_m,\imath_m]\leq[M_t,g_t,\imath_t]$. On the other hand, since $(M_t,g_t,\imath_t)$ is a solution and $(M_m,g_m,\imath_m)$ is maximal also the converse holds, so $(M_m,g_m,\imath_m)\cong(M_t,g_t,\imath_t)$. □

Adding regularity conditions on the extensions,[528] this would be a meaningful and natural PDE version of strong cosmic censorship. Indeed, the requirement that also the (strongly censored) extension satisfies the vacuum Einstein equations provides these regularity conditions, at least up to a point: one may either go for extensions in which the metric is C^2, i.e. the borderline case where Einstein equations make sense classically, or allow C^0 metrics as long as the associated Christoffel symbols are locally L^2, which is the least regular case in which the metric can still be defined as a weak solution to Einstein's equations.[529] Indeed, a weak solution of the vacuum Einstein equations is a metric g for which for all compactly supported $X,Y\in\mathfrak{X}(M)$,

$$\int_M d^4x\sqrt{-\det(g(x))}R_{\mu\nu}(x)X^\mu(x)Y^\nu(x)=0. \tag{10.84}$$

Partial integration shows that this is well defined iff the $\Gamma^\rho_{\mu\nu}$ are locally L^2.

However, in the following PDE definition of strong cosmic censorship the extension is not required to satisfy the Einstein equations! For convenience also state the weak PDE version.

Definition 10.13

- *The* PDE-**strong cosmic censorship conjecture** *states that the* MGHD *of "generic" complete initial data is inextendible (in a regularity class to be specified).*
- *The* PDE-**weak cosmic censorship conjecture** *states that if "generic" complete initial data have a* MGHD *that is asymptotically flat at null infinity (and hence admits a conformal completion to begin with), then future null infinity $\mathscr{I}^+$ of this* MGHD *is complete.*[530]

[528] Chruściel, Isenberg, & Moncrief (1990) and Chruściel & Isenberg (1993) consider smooth extensions.

[529] See Geroch & Traschen (1987), Christodoulou, (2009), p. 9, and Luk (2017), footnote 1. This simple observation should not be confused with the very deep result that having the *Ricci tensor* in L^2 is sufficient for the (vacuum) Einstein equations to be weakly *solvable* at least locally (Klainerman, Rodnianski, & Szeftel, 2015).

[530] See Definition 10.2.3. Christodoulou (1999a) reformulates this definition of weak cosmic censorship in such a way that the idealization $\mathscr{I}^+$ no longer occurs. Let (Σ,h,K) be asymptotically flat initial data for the Einstein equations (satisfying the constraints), with MGHD (M,g,i). Christodoulou (1999a) then defines (M,g) to have "complete future null infinity" iff for any $s>0$ there exists a region $B_0\subset B\subset\Sigma$ such that $\partial D^+(B)$, which is ruled by lightlike geodesics, has the property that each lightlike geodesic starting in $\partial J^+(B_0)\cap\partial D^+(B)$ can be future extended beyond parameter value s. Here $D^+(B)$ is the future domain of dependence of B, and each lightlike geodesic in question is supposed to have tangent vector $L=T-N$, where T is the fd unit normal to Σ in M and N is the outward unit normal to ∂B in Σ. See Christodoulou & Klainerman (1993) for background on these constructions.

In view of the path from Penrose to PDE described above, part 1 of this definition is stronger than the application of Penrose's definition to the maximal development of the given initial data in the sense of Definition 10.11. As a compromise, one might pose the conjecture relative to extensions that satisfy some curvature condition, such as the timelike and/or null curvature conditions assumed in the singularity theorems of Hawking and/or Penrose, respectively.

Although it would make sense in general, in practice the strong conjecture is posed for either non-compact Cauchy surfaces Σ with asymptotically flat initial data, or for compact Σ (also called the 'cosmological' case).[531] Except in relatively simple cases like Minkowski space-time, in the asymptotically flat case the validity of PDE-strong cosmic censorship turns out to be very sensitive to the precise regularity of the extension.[532] Sensitivity to the precise formulation of the genericity conditions has not been much discussed in the literature,[533] but already the simplest examples (see §10.6) show that such conditions are necessary: strong cosmic censorship in any version already fails for all parameter values of the Kerr metric (as lone as $a \neq 0$ and $m \neq 0$), and even for the Reissner–Nordström metric (again with $e \neq 0$ and $m \neq 0$); in the exact black hole solutions discussed in this book it only holds (in both versions) for the Kruskal space-time.

This made it especially courageous–or some might say reckless–of Penrose to formulate the strong conjecture. But of course he would not have done so without good arguments against the counterexamples. His key observation, one of his most prophetic insights, was first published in 1968 (i.e. before even the weak version, which predated the strong one, was formulated), viz.[534]

> There is a further difficulty confronting our observer who tries to cross H_C^+. As he looks out at the universe he is "leaving behind," he sees, in one final flash, as he crosses H_C^+, the entire later history of the rest of his "old universe." If, for example, an unlimited amount of matter eventually falls into the star then presumably he will be confronted with an infinite density of matter along "H_C^+". Even if only a finite amount of matter falls in, it may not be possible in generic situations to avoid a curvature singularity in place of H_C^+. This is at present an open question. But it may be, that the place to look for curvature singularities is in this region rather than (or as well as?) at the "center." (Penrose, 1968, p. 222)

> Our contention in this note is that if the initial data is generically perturbed then the Cauchy horizon does not survive as a non-singular hypersurface. It is strongly implied that instead, genuine space-time singularities will appear along the region which would otherwise have been the Cauchy horizon. (Simpson & Penrose, 1973, p. 184)

Note that from Penrose's point of view H_C^+ is a (future) Cauchy horizon relative to some wannabe Cauchy surface Σ inside some "large" analytically (possibly maximally) extended space-time.

[531] See Ringström (2009) and Doboszewski (2017) for reviews of the cosmological case, where the strong (PDE) conjecture seems to hold "generically" in regularity C^2 (and of course higher).

[532] Proposition 6.2 is concerned with smooth extensions, but the version in Chruściel (2020), i.e. Proposition 4.4.3, originating in Chruściel & Costa (2008), even works for C^k extensions with $k \geq 2$. Hence a proof of inextendibility based on an explicit classification of all causal or even just all timelike geodesics, works for any $k \geq 2$. The inextendibility of both Minkowski space-time and Kruskal space-time can be proved in this way; see e.g. Corollary 13.37 in O'Neill (1983). Minkowski space-time even turns out to be inextendible in C^0, which is a far more difficult result (Sbierski, 2018ab). So here the validity of PDE-strong cosmic censorship is independent of the regularity. However, for two-ended asymptotically flat data for the spherically symmetric Einstein–Maxwell-scalar field system (to which the conjecture, so far discussed for the vacuum case, can be extended in the obvious way), the conjecture *fails* in C^0, i.e. the MGHD is extendible with a C^0 metric, but it *holds* in C^1, in that the metric of the extension fails to be C^1 (Dafermos, 2003, 2005). The situation for the Kerr metric is similar (Dafermos & Luk, 2017).

[533] Specific papers that clearly state such conditions include Dafermos (2003) and Luk & Oh (2019a), §3.

[534] Here H_C^+ is the Cauchy horizon, which in the original text is denoted by $H_+(\mathscr{H})$. This is the only change.

From the PDE point of view, on the other hand, H_C^+ is the boundary of the MGHD of the corresponding initial data on Σ. Either way, this "blueshift instability" of H_C^+ has been confirmed in a large number of studies and hence remains the key to proofs of PDE-strong cosmic censorship.[535]

Failure of strong cosmic censorship is often taken to imply a failure of determinism in GR.[536] This is true in a specific sense, which is slightly different for the Penrosian and the PDE versions, but in both cases rests on the idea that the (classical) world–including the gravitational field itself, or at least its physical degrees of freedom–is governed by hyperbolic partial differential equations whose initial data should be given on a hypersurface Σ and whose solutions should be thereby *determined* on its domain of dependence $D(\Sigma)$.[537] If our space-time (M,g) is globally hyperbolic, then it has a Cauchy surface Σ such that $D(\Sigma) = M$ (see Proposition 5.38), and hence all (scientific) things in M are determined by their values on Σ. In PDE language, this formalizes the notion of ***Laplacian determinism***,[538] which in fact goes back (at least) to Leibniz:

> One sees then that everything proceeds mathematically - that is, infallibly - in the whole wide world, so that if someone could have sufficient insight into the inner parts of things, and in addition has remembrance and intelligence enough to consider all the circumstances and to take them into account, he would be a prophet and would see the future in the present as in a mirror.[539] (Leibniz, undated)

> An intelligence which could comprehend all the forces that set nature in motion, and all positions of all items of which nature is composed–an intelligence sufficiently vast to submit these data to analysis–it would embrace in the same formula the movements of the greatest bodies in the universe and those of the lightest atom; for it, nothing would be uncertain and the future, as well as the past, would be present to its eyes.[540] (Laplace, 1814)

From Penrose's point of view, if a space-time (M,g) fails to be globally hyperbolic, then, taking some wannabe Cauchy surface Σ, neither the part $M\backslash D(\Sigma) \neq \emptyset$ of space-time behind the Cauchy horizon of Σ itself, nor things happening behind this horizon, are determined by initial data on Σ.

[535] In the wake of Penrose (1968), see Simpson & Penrose (1973), McNamara (1978), Hiscock (1981), down to Chesler, Narayan & Curiel (2020). Mathematically rigorous work started with Dafermos (2003); more recent papers may be traced back from Van de Moortel (2020). The conclusion seems to be that Cauchy horizons turn into so-called ***weak null singularities***, which are null boundaries with C^0 metric but $\Gamma^\rho_{\mu\nu}$ not locally L^2 (Luk & Sbierski, 2016; Luk, 2017). At least for one-ended asymptotically flat initial data, behind such a weak mull singularity there is also a strong curvature singularity at $r = 0$. See also Luk & Oh (2019ab) for the two-ended case, and Gajic & Luk (2017) for extremal Reissner–Nordström black holes (i.e. $|e| = m > 0$). Such results concern cosmological constant $\lambda = 0$. See Dias *et al.* (2018ab) and Luna *et al.* for $\lambda > 0$ (specifically de Sitter space), whose verdict on strong cosmic censorship depends critically on both the matter coupling and the regularity of the extension. For $\lambda < 0$ even weak cosmic censorship seems to fail (Hertog, Horowitz, & Maeda, 2004; Crisford & Santos, 2017).

[536] Earman (1995) is a classic, taken up among others by Doboszewski (2017, 2019, 2020) and Manchak (2020).

[537] The classical exposition of this world view is Courant & Hilbert (1962), which unfortunately does not cover GR. In that light, see also Choquet-Bruhat (2009) and, specifically for matter fields, Bär, Ginoux, and Pfäffle (2007).

[538] Succinctly: 'The world W is Laplacian deterministic just in case for any physically possible world W', if W and W' agree at any time, then they agree at all time.' (Earman, 1986, p. 13). Hence we assume Σ to be spacelike.

[539] The undated German original is quoted by Cassirer (1936), pp. 19–20: 'Hieraus sieht man nun, das alles mathematisch, d.i. uhnfehlbar zugehe in der ganzen weiten Welt, so gar, dass wenn einer eine genugsame Insicht in die inneren Teile der Dinge haben könnte, und dabei Gedächtnis und Verstand genug hätte, um alle Umstände vor zu nehmen und in Rechnung zu bringen, würde er ein Prophet sein, und in dem Gegenwärtigen das Zukünftige sehen, gleichsam als in einem Spiegel.' English translation by the author.

[540] Translation taken from the English edition from 1902, p. 4. Note that Leibniz' prophet appeals to the logical structure of the universe that makes it deterministic, whereas Laplace's intelligence knows (Newtonian) physics. Van Strien (2014) argues that Laplace also falls back on Leibniz and (uncharacteristically) gets the physics wrong by not mentioning the momenta that the intelligence should know, too, besides the forces and positions.

From the PDE point of view, although any MGHD (M',g') is globally hyperbolic with Cauchy surface $\Sigma \subset M'$, if (M',g') is extendible with extension (M,g), then Σ fails to be a Cauchy surface for $M \supset M'$ and we are back to Penrose's perspective; this is true even if the extension (M,g) satisfies the Einstein equations. A difference between Penrose and PDE arises if the extension of "the" MGHD is not unique, which happens in some examples.[541] In that case the lack of *apparent* determination is even worse, but otherwise the analysis is not greatly affected.

Either way, a lack of global hyperbolicity of (M,g) does not imply that space-time is *in*deterministic in the sense that random events occur in $M\backslash D(\Sigma)$, as in quantum mechanics.[542] The point is rather that events beyond the Cauchy horizon of Σ are not determined by the initial data originally expected to do so. This is quite remarkable, but nonetheless such events may instead be determined by signals coming from a (locally) naked singularity, or, should it turn into some kind of weak singularity itself, as mentioned above, by events happening on the Cauchy horizon. To further weaken the connection between global hyperbolicity and determinism, let us note that the undeniable indeterminism of someone falling into a black hole singularity is perfectly well compatible with global hyperbolicity, as the Schwarzschild solution shows.

More generally, in classical (mathematical) physics indeterminism may come from either a lack of *uniqueness* of solutions or from a lack of *existence* thereof; the latter includes incompleteness of solutions, i.e. non-existence after (or before) some finite time. Strong cosmic censorship tries to secure *uniqueness* at the level of the Einstein equations (where *existence* is secure), but as we have seen its failure does not automatically imply indeterminism *per se*. In our view, lack of existence (e.g. for geodesic equations) is the more relevant source of indeterminism in GR.[543]

We briefly return to weak cosmic censorship, where the connection between the Penrosian and the PDE versions is less clear than for strong cosmic censorship. Definition 10.13 is identical to clause 3 in Definition 10.2. Although both contexts use (future) null infinity $\mathscr{I}^+$, the connection with singularities/incompleteness, which indeed should be irrelevant to the concept of $\mathscr{I}^+$, seems missing in connection with weak cosmic censorship.[544] Also, whereas Penrose's version states that *outgoing* signals from a black hole singularity are blocked by an event horizon H_E^+, the PDE version is about *incoming* signals: the further these are away from H_E^+, the longer it takes them to enter H_E^+, and in the limit at null infinity this takes infinitely long. Nonetheless, in §10.6 we shall see that in simple examples they match, because lack of global hyperbolicity of $I^-(\mathscr{I}^+)$ gives a wannabe Cauchy surface Σ a Cauchy horizon which cuts off $\mathscr{I}^+$.

In sum, there is no unique concept of weak or strong cosmic censorship. As a compromise, one might summarize the conjectures as follows. In "physically reasonable" space-times:[545]

- weak cosmic censorship postulates the *appearance and stability of event horizons*;[546]
- strong cosmic censorship requires the *instability and disappearance of Cauchy horizons*.

[541] This is especially true in the cosmological case. Examples include Misner space-time, Taub–NUT space-time, polarized Gowdy space-times, etc. See Earman (1995), Ringström (2009, 2010), and Doboszewski (2017).

[542] The proof of this shocking claim, going back to Born (1926), is in fact very recent (Landsman, 2020, 2021).

[543] In *non-relativistic* mechanics bodies may disappear to infinity in finite time (Xia, 1992; Saari & Xia, 1995), and hence, by the same (time-reversed) token, may *appear* from nowhere in finite time and hence influence affairs in a way unforeseeable from any Cauchy surface. This analogy with GR is discussed by Earman (2007), §3.6.

[544] After a talk by the author on June 16, 2021, Mihalis Dafermos pointed out that PDE-weak cosmic censorship should be seen as "weak weak cosmic censorship", which is something like a test case for Definition 10.8.2. Yet neither Penrosian weak cosmic censorship nor PDE-strong cosmic censorship implies PDE-weak cosmic censorship.

[545] Penrose's "physically reasonable" is preferable to the mathematicians' "generic", since the so-called ***fine-tuning problem*** suggests that *our cosmos is not al all generic!* See Landsman (2016) and Adams (2019) for introductions.

[546] Whenever, of course, these are expected, viz. when trapped surfaces form in gravitational collapse (Joshi, 2007).

10.6 Cosmic censorship in some simple examples

In this section we analyze the relationship between the Penrosian and the PDE versions of the cosmic censorship from three key black hole examples and their Penrose diagrams:[547]

- Maximally extended Schwarzschild (i.e. Kruskal) with $m > 0$ (and two-sided initial data);
- Schwarzschild with $m < 0$, whose singularities and horizons looks like supercharged Reissner–Nordström ($|e| > m > 0$), or ultrafast rotating Kerr ($|a| > m > 0$);
- Reissner–Nordström with $0 < |e| < m$, which also resembles Kerr with $0 < |a| < m$.

In the first case the solution coincides with the MGHD of the pertinent (two-ended) initial data, so the difference between strong Penrosian and strong PDE cosmic censorship fades. We have already drawn the Penrose diagram of the maximally extended Schwarzschild solution with $m > 0$ in §10.3. The maximal Cauchy development of a generic two-sided Cauchy surface Σ with suitable initial data (drawn as a horizontal blue line) is simply the entire space-time. In particular, the Cauchy horizon $H_C^{\pm}$ of Σ is empty. The upper two green lines form the future event horizon H_E^+ of the black hole area, which is the upside-down upper triangle (labeled region II), whereas the lower two green lines form the past event horizon H_E^- of the white hole area, i.e. the lower triangle (region IV). The right-hand diamond is region I, the left-hand diamond is region III. Fd causal curves cannot *leave* region II and they cannot *enter* IV.

Both cosmic censorship conjectures hold in both versions (i.e. Penrose and PDE):[548]

- *Weak cosmic censorship for Kruskal space-time.*

  Penrose: Σ is a Cauchy surface for $I^-(\mathscr{I}^+)$, making it globally hyperbolic.[549]

  PDE: each component of $\mathscr{I}^+$ ends at timelike infinity i^+ and hence its lightlike geodesics are future complete (as confirmed by parametrization and computation).

- *Strong cosmic censorship for Kruskal space-time.*

  Penrose: Kruskal space-time is globally hyperbolic (since the causal structure of the diagram is such that the line Σ represents a Cauchy surface).

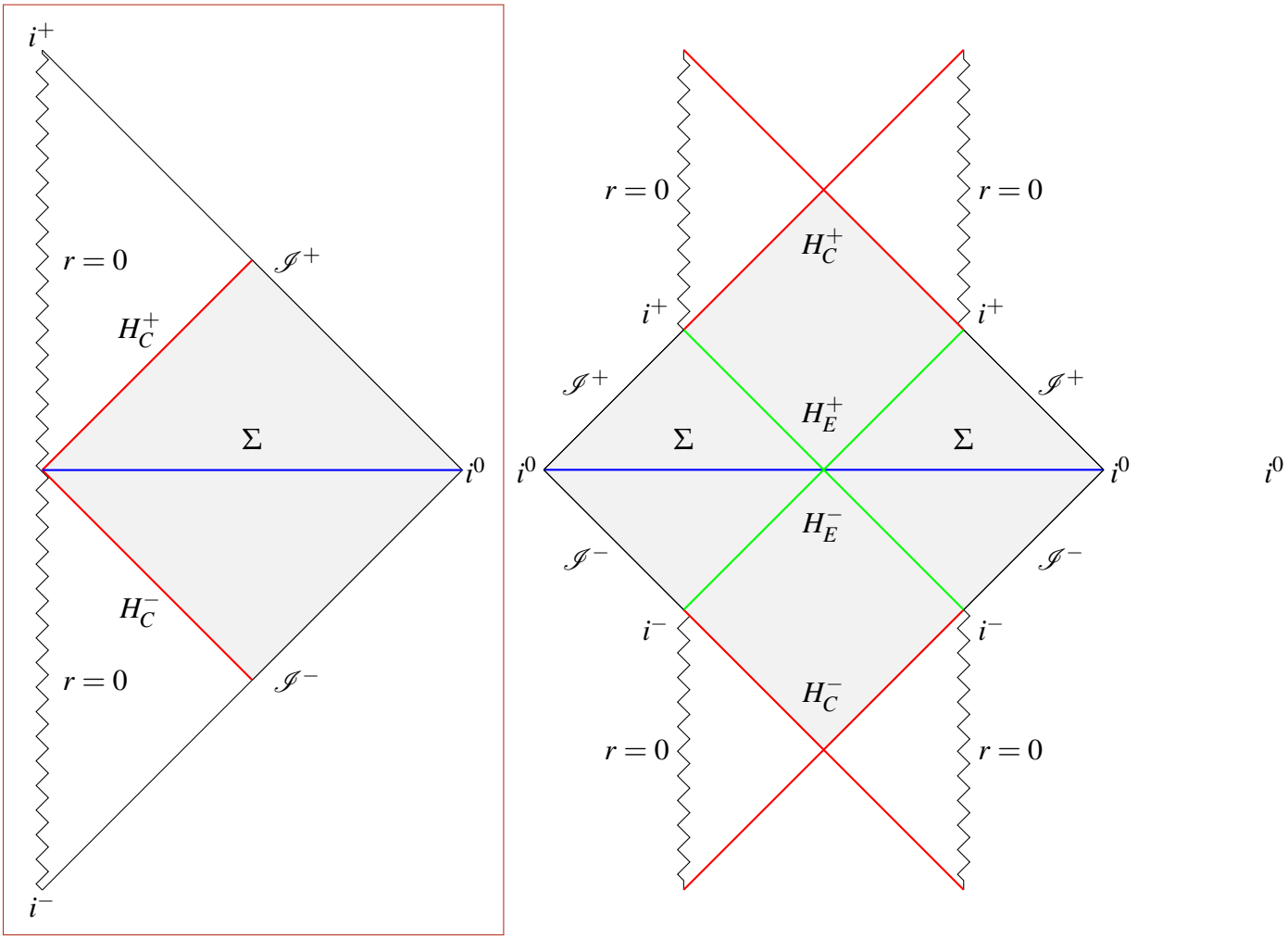
  PDE: Explicit classification of the causal geodesics in Kruskal space-time (M_K, g_K) shows that the antecedent of the second ("or") part of Proposition 6.2 is satisfied: a causal geodesic is incomplete iff it crashes into the singularity at $r = 0$, in which case it has unbounded curvature because of (9.18). Otherwise, it goes to infinity, in which case it is complete. Hence Kruskal space-time is inextendible (cf. footnote 532).

However, for $m < 0$ Kruskal, Reissner–Nordström, and Kerr, differences arise between the Penrosian and the PDE perspectives, since in these cases the maximal (analytic) solutions differ from the MGHD of the pertinent initial data. In particular, although (curvature) singularities are not part of space-time in any case, they can at least be drawn as boundaries in the maximal solutions, where they lie behind a Cauchy horizon. But precisely for that reason singularities are beyond the scope of the corresponding MGHD. Here are the Penrose diagrams:

[547] This section is largely based on Hawking & Ellis (1973), pages 158 and 160, as well as on Dafermos & Rodnianski (2008) and Dafermos (2014ab, 2017, 2019) for the PDE side.

[548] In view of the recently proved stability of Schwarzschild space-time (Dafermos *et al.*, 2021), they also hold in the informal version stated at the end of the previous section. So far, such a proof is lacking for the other cases.

[549] Alternatively: any *incomplete* future inextendible timelike curve c must crash in the upper $r = 0$ singularity. Hence $I^-(c)$ lies partly in region II, which is disjoint from $I^-(\mathscr{I}^+)$, so that $I^-(c) \not\subseteq I^-(x)$ for all $x \in I^-(\mathscr{I}^+)$.



Left: *Penrose diagram of $m<0$ Schwarzschild, or supercharged Reissner–Nordström ($|e|>m>0$), or fast Kerr ($|a|>m>0$). These solutions have a singularity at $r=0$, but unlike the $m>0$ Kruskal case it is not shielded by an event horizon. The red lines labeled H_C^- and H_C^+ are past and future Cauchy horizons with respect to the blue line, which indicates a maximal spacelike surface whose initial data give rise to the metrics in question and whose maximal Cauchy development* (MGHD) *is the grey area.*

Right: *Penrose diagram of subcritical Reissner–Nordström ($0<|e|<m$), whose event and Cauchy horizons (despite the different structure of the singularity) also resemble those of slowly rotating Kerr ($0<|a|<m$). The maximal Cauchy development* (MGHD) *of the pertinent initial data given on the maximal spacelike hypersurface represented by the blue line labeled Σ is again colored in grey. It contains past and future event horizons labeled H_E^- and H_E^+, drawn in green, but unlike the $m>0$ Schwarzschild case, the singularity they are supposed to shield cannot be reached directly from the maximal Cauchy development, which is bounded by the various fictitious boundaries $\mathscr{I}^\pm$, $i^\pm$, and i^0, which lie at infinity, as well as by the Cauchy horizons $H_C^\pm$, drawn in red, which can be reached in finite proper time.*[550]

Despite the different space-times they apply to, the outcomes of the Penrosian version and the PDE version of both weak and strong cosmic censorship are once again the same:[551]

[550] This diagram can be infinitely extended in both directions (Hawking & Ellis, 1973, pp. 158, 165): to the north, another grey area folds inside the upper two red line segments, and similarly to the south, but we do *not* do so here.

[551] For $m<0$ Kruskal the initial data are not complete in this case, so strictly speaking the cosmic censorship conjectures do not apply here. Nonetheless, they can be stated and the comparison is instructive.

- $m < 0$ Kruskal (etc.): For the Penrosian total space-time the difference between weak and strong cosmic censorship evaporates, since $I^{-1}(\mathscr{I}^+) = M$, which is not globally hyperbolic: wherever one tries to place a wannabe Cauchy surface Σ (such as the blue line), above the surface inextendible causal curves can be drawn that enter i^+ or $\mathscr{I}^+$ in the future and enter the singularity at $r = 0$ in the past, without crossing Σ. Similarly, below Σ one may draw inextendible causal curves converging to the singularity in the future, and to i^- or $\mathscr{I}^-$ in the past, which once again do not cross Σ. Thus neither weak nor strong cosmic censorship holds for this space-time.

  The PDE picture applies to the grey area, which is the MGHD of the initial data given on the blue line marked Σ in the left-hand Penrose diagram. Then weak cosmic censorship fails because future null infinity $\mathscr{I}^+$ is clearly incomplete: lightlike geodesics terminate at the Cauchy horizon (where they "fall off' space-time) and hence are incomplete. On the other hand, strong cosmic censorship fails because the grey space-time, though globally hyperbolic (in contrast with the entire space as we have just seen), is evidently (smoothly–even analytically) extendible, namely by the total space displayed. Though they do not coincide, we see that strong and weak cosmic censorship are closely related: future incompleteness of lightlike geodesics at null infinity happens because the MGHD is extendible.

- Subcritical Reissner–Nordström ($0 < |e| < m$): for both Penrose and PDE strong cosmic censorship fails, whereas the weak version holds. In the Penrosian version the total space fails to be globally hyperbolic because of the part above the grey area (i.e. beyond the future Cauchy horizon H_C^+): one has past-directed inextendible causal curves that (backwards in time) end up in the singularity and hence never cross Σ (e.g. those crossing the upper left, NW-pointing red line from N to SW). Weak cosmic censorship holds because of the future event horizon H_E^+, which shields the upper $r = 0$ singularity above it. Equivalently, $I^-(\mathscr{I}^+)$ is globally hyperbolic, a property it inherits from the MGHD.[552]

  The PDE view is cleaner here: roughly speaking, as in the $m > 0$ Kruskal or Schwarzschild case (but unlike the $m < 0$ case) future null infinity $\mathscr{I}^+$ ends at future timelike infinity i^+ and hence is complete, so that weak cosmic censorship holds.[553] Strong cosmic censorship, on the other hand, fails because the MGHD (marked in grey) is clearly smoothly extendible, namely into, for example, the space-time shown.

If the strong Penrosian conjecture fails for some space-time (M_P, g_P), then its lack of global hyperbolicity typically occurs because (M_P, g_P) is an extension of the MGHD (M, g) of some given initial data, whose Cauchy surface Σ fails to be one for (M_P, g_P). Similarly, if $I^{-1}(\mathscr{I}^+)$ is not globally hyperbolic (so that there is a naked singularity), M_P usually comes from extending some (M, g), as above, whose Cauchy surface becomes a wannabe Cauchy surface in M_P, with an associated future Cauchy horizon that cuts off $\mathscr{I}^+ \cap \tilde{M}$, causing its incompleteness.[554]

As already mentioned, the fact that these well-known examples *violate* (at least) strong cosmic censorship makes it all the more remarkable that the ensuing conjecture was made in the first place. In order to save it, such examples must be shown to be "non-generic", for example through the blueshift instability mentioned in the previous section, or some other mechanism.

[552] This is no longer true for the maximal extension, which adds countably many components of $\mathscr{I}^+$. Keeping the single Σ shown would allow many causal curves in $I^-(\mathscr{I}^+)$ not hitting it, but adding countably many copies of Σ in the obvious way would allow causal curves hitting this total Σ many times, which then cannot be a Cauchy surface.

[553] It even holds in the maximal extension, driving the Penrose and PDE versions apart!

[554] However, these aren't rigorous deductions: there are pathological cases where strong cosmic censorship holds whilst the weak version fails. See e.g. the Penrose diagram at the end of §2.6.2 of Dafermos & Rodnianski (2008).

10.7 Structure of event horizons and Cauchy horizons

Most of the physics of black holes, including cosmic censorship, is concerned with various kinds of horizons. Three important types of black hole horizons one needs to be familiar with are:[555]

- ***Event horizons***, defined in (10.79) based on Penrose's concept of null infinity $\mathscr{I}$, i.e.

$$H_E^{\pm} = \partial I^{\mp}(\mathscr{I}^{\pm}); \tag{10.85}$$

- ***Cauchy horizons***, defined in (5.178) - (5.182), applied to wannabe Cauchy surfaces S, i.e.

$$H_C^{\pm}(S) = \partial D^{\pm}(S) \backslash S. \tag{10.86}$$

- ***Killing horizons*** (for stationary black holes), still to be defined, see §10.8.

These are all null hypersurfaces, as we will now prove for the first two cases (for the third it will be true by definition). For convenience, let us recall some relevant definitions from chapter 5:

Definition 10.14

- *A subset $S \subset M$ is* **acausal** *if no* causal *curve starts and ends at S.*
- *A subset $S \subset M$ is* **achronal** *if no* timelike *curve starts and ends at S.*
- *The* **edge** *of an* achronal *set S consists of all $x \in M$ for which every open nbhd U of x contains points y and z and two timelike curves from y to z, of which just one intersects S.*
- *A* **future/past set** *set $F \subset M$ satisfies $I^{+/-}(F) \subset F$ (if F is open this implies $I^{\pm}(F) = F$).*
- *An* **achronal boundary** *is a set ∂F where F is a future set.*[556]
- *The* **domain of dependence/influence** *$D^{+/-}(S)$ of $S \subset M$ is the set of all $x \in M$ for which every past/future-inextendible pd/fd causal curve starting from x intersects S.*
- *The* **domain of dependence** *of S is the union $D(S) = D^+(S) \cup D^-(S)$.*
- *A* **wannabe Cauchy surface** *is an* acausal edgeless *(and hence closed) subset of M.*
- *The* **future/past Cauchy horizon** of a wannabe Cauchy surface S *is given by* (10.86).
- *The* **Cauchy horizon** of a wannabe Cauchy surface S *is $H_C(S) = \partial D(S)$.*
- *A* **Cauchy surface** *is a wannabe Cauchy surface S for which $D(S) = M$, i.e. $H_C(S) = \emptyset$.*

Further to Lemma 5.37, we collect some of the properties of such sets, without proof:[557]

[555] Apparent horizons are briefly discussed in §10.11. See also footnote 508.

[556] For $F = I^+(A)$, below (5.146) we already showed that an achronal boundary is indeed achronal. In general, if $y \in I^+(x)$ for $x, y \in \partial F$, then $y \in I^+(\overline{F}) = I^+(F)$. But this is open, so $y \in \text{int}(F)$ whilst $y \in \partial F$, which is a contradiction. Conversely, a *maximal* achronal set is an achronal boundary, and since any achronal set is contained in a maximal one, any achronal set is contained in an achronal boundary. See Minguzzi (2019), Theorem 2.87.

[557] No. 1 is Proposition 2.136 in Minguzzi (2019), and the case $F = I^+(A)$ is Claim 2 on page 12 of Galloway (2014). No. 2 is trivial, since $I^+(I^+(A)) = I^+(A)$ and $I^+(A)$ is open. No. 3 is Proposition 3.15 in Penrose (1972).

Lemma 10.15 *1. Achronal boundaries are edgeless.*

2. Any $F = I^{+/-}(A)$ is an open future/past set, for arbitrary $A \subset M$.

3. Given an achronal boundary $B = \partial F$, where F is a future set, there is a unique disjoint decomposition $M = P \cup B \cup F$, where P is a past set and $B = \partial P$ (and likewise $F \leftrightarrow P$).[558]

It follows that both event horizons and Cauchy horizons of wannabe Cauchy surfaces are closed edgeless achronal topological hypersurfaces. But so are spacelike Cauchy surfaces in globally hyperbolic space-times, so our horizons must have special features that make them contain sufficiently many causal curves so as to become lightlike according to Corollary 5.15. These features are different, since although according to (10.85) - (10.86) both horizons are (part of) topological boundaries, future or past sets are very different from domains of dependence. Yet the second case will be reduced to the first! The key proposition for this is as follows:[559]

Proposition 10.16 *1. Let $S \subset M$ be a closed subset of M with associated achronal boundary*

$$B = \partial I^+(S). \tag{10.87}$$

If $x \in B \backslash S$, there exists a fd lightlike geodesic γ that is contained in B with future endpoint x and either a past endpoint on S or no past endpoint at all (i.e. γ is past inextendible).

2. Let $S \subset M$ be a closed achronal *subset of M with associated future Cauchy horizon $H^+_C(S)$. If $x \in S \backslash \mathrm{edge}(S)$, there exists a fd lightlike geodesic γ that is contained in $H^+_C(S)$ with future endpoint x and either a past endpoint on* $\mathrm{edge}(S)$ *or no past endpoint at all (*ibid.*).*

For an example of the first case of part 1, take $S = \{0\}$, where 0 is the origin in $\mathbb{M}$. Then B is the closed forward lightcone emanating from the origin, which includes the origin. For the second case, consider the left-hand figure in the next section §10.8, and take S to be the left-most accelerated curve in region I. Then B is the entire SE–NW axis (i.e. $x = -t$) and so no pd lightlike geodesic ever touches S. Both cases of part 2 can be covered by a single example, see next page.

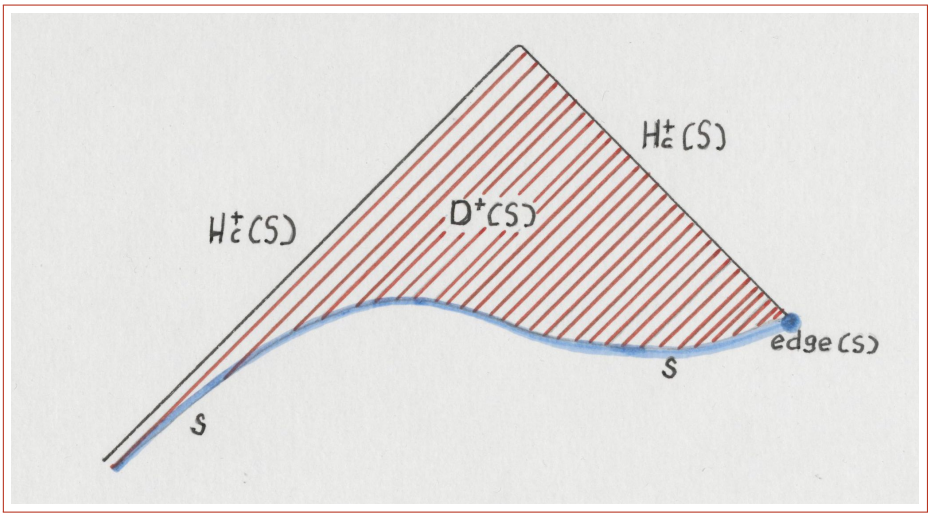
A similar result holds with past and future interchanged. If we add this, and in part 1 take $S = \mathscr{I}^{\pm}$, noting that the boundaries $\mathscr{I}^{\pm}$ are closed in $\hat{M}$ (as $i(M) \subset \hat{M}$ is open by construction), we obtain a result about event horizons. If in part 2 we take S to be a wannabe Cauchy surface and note that $\mathrm{edge}(S)$ is empty in that case, we have a result about Cauchy horizons. Thus:

Corollary 10.17 *1. Let $H^{+/-}_E$ be the future/past event horizon of a black/white hole. Then any $x \in H^{+/-}_E$ lies on a future/past intextendible lightlike geodesic contained in $H^{+/-}_E$.*

2. If $H^{+/-}_C(S)$ is the future/past Cauchy horizon of a wannabe Cauchy surface S, any $x \in H^{+/-}_C(S)$ lies on a past/future intextendible lightlike geodesic contained in $H^{+/-}_C(S)$.

[558] Penrose (1972), Remark 3.16, warns that although in Minkowski space one has $F = I^+(B)$ and $P = I^-(B)$ this need not be true in general, with a specific counterexample already in $(0,1) \times \mathbb{R} \subset \mathbb{M}_2$ with Minkowski metric.

[559] Both results are are due to Penrose (1972), Theorems 3.20 and 5.12, though neither is stated in the context of black holes! The first one may also be found in e.g. Wald (1984), Theorem 8.1.6, Galloway (2014), Proposition 3.4, and Minguzzi (2019), Lemma 2.89 and Corollary 2.92. The second is Wald (1984), Theorem 8.3.5, Galloway (2014), Proposition 5.3, and Minguzzi (2019), Theorem 3.24. Each author uses a slightly different version of the curve limit lemma. Perhaps Penrose's original proofs are now seen as heuristic, but in our view they are very clear.



A closed achronal subset S of 2d Minkowski space-time, drawn in blue, starts at edge(S) *on the right, and then, always staying spacelike, asymptotes to the left-hand side of the backward lightcone off the origin. Its domain of dependence $D^+(S)$ is drawn in red and its future Cauchy horizon $H^+_C(S)$ consists of the two black lines, of which the right one ends at and includes* edge(S)*, whilst the left one goes on downward forever along the lightcone. Past-directed lightlike geodesics within the right-hand branch of $H^+_C(S)$ end at* edge(S) *(after which they may eave $H^+_C(S)$), whereas those on the left are past inextendible.*[560]

Thus $H^{+/-}_E$ is ruled by future/past intextendible lightlike geodesics (called the ***generators***) of $H^{+/-}_E$),[561] and similarly $H^{+/-}_C(S)$ is ruled by past/future intextendible lightlike geodesics. Hence both event horizons $H^{\pm}_E$ and Cauchy horizons $H^{\pm}_C(S)$ are (topological) null hypersurfaces.

We now prove case 1 of Proposition 10.16 for the special case $S = \{y\}$, so that $B = \partial I^+(y)$. This proof contains the idea of the general case. So take $x \in \partial I^+(y)$, then by definition there is a sequence (x_n) in $I^+(y)$ converging to x, and there are pd timelike curves γ_n from x_n to y, with

$$\gamma_n : [0, b_n] \to M; \qquad \gamma_n(0) = x_n; \qquad \gamma_n(b_n) = y, \tag{10.88}$$

parametrized by h-arc length, cf. §5.6. By the curve limit lemma 5.26, there is a limit curve

$$\gamma : [0, b] \to M; \qquad \gamma_n(0) = x; \qquad \gamma(b) = y, \tag{10.89}$$

where $b_n \to b$. This limit curve is *causal* and, coming from curves γ_n in $I^+(y)$ as a uniform limit, it lies in the closure $\overline{\partial I^+(y)}$, which consists of $\partial I^+(y)$ and its boundary B. If γ contained any point $z \in I^+(y)$, then $x \in I^+(y)$ by Proposition 5.4.5, but $I^+(y)$ is open and $x \in \partial I^+(y)$, so this is impossible. Hence γ must lie entirely in the achronal set B, so that by Corollary 5.15 it must be a lightlike (pre)geodesic (which after reparametrization, if necessary, becomes a geodesic). Finally, if γ has a past endpoint w in B different from y, then the above construction could be repeated with x in the role of w, duly extending γ. See footnote 563 for more information.
For the case of a general closed set S, we must replace the third entry in (10.88) by

$$\gamma_n(b_n) = y_n; \qquad (y_n \in S). \tag{10.90}$$

[560] Figure redrawn (and adapted) from Penrose (1972), Fig. 36, by Edith de Jong.

[561] These lightlike geodesics are intextendible *in* M; by Proposition 10.16 they may have future/past endpoints *in* $\mathscr{I}^{+/-}$, as well as past/future endpoints in $H^{+/-}_E$ itself, before/after which they leave the horizon. This leaves the possibility that lightlike geodesic on $H^{\pm}_E$ hit a singularity. In the proof of Hawking's Area theorem (see §10.12) this is excluded by some version of weak cosmic censorship. See also Wald (1994), §6.1.

If the sequence (y_n) has an accumulation point y the proof is almost the same as above, since the limit curve satisfies (10.89). If not, we use a trick:[562] take a (geodesically) convex nbhd U of x with compact closure and hence compact boundary ∂N, and take the intersections $z_n = \gamma_n \cap U$. By compactness of ∂N the z_n have an accumulation point z. We now restart the argument with y_n replaced by z_n, which works because $z_n \in I^-(x)$. This gives a causal limit curve from x to z, which by the above reasoning must lie in B and must be a lightlike geodesic within B, etc.[563]

We now prove part 2 of Proposition 10.16 by reduction to case 1, which is possible because $H_C^\pm$ turns out to part of an achronal boundary.[564] Indeed, if, for $H_C^+(S)$ to be concrete, we define

$$W := I^+(H_C^+(S)) = I^+(S) \backslash \overline{D^+(S)}, \tag{10.91}$$

where either side could be taken as the definition and the other as an inference, we have

$$H_C^+(S) = \partial W \cap \overline{D^+(S)}; \qquad \partial W = H_C^+(S) \cup \partial I^+(S) \backslash S, \tag{10.92}$$

and similarly for the past Cauchy horizon $H_C^-(S)$. This is easily proved, and verified in the picture above. Let us also give two examples where S is edgeless (see next page):

- The upper picture is $2d$ Minkowski space-time with $(1,1)$ deleted, and the x-axis is taken to be our wannabe Cauchy surface S (which by definition is acausal and edgeless).

- In the Quinten space-time $\mathbb{M}_2'$ (i.e. $\mathbb{M}_2$ with the closed horizontal line segment from $(t,x) = (2,-1)$ to $(2,1)$ removed), our wannabe Cauchy surface S is again the x-axis. Unlike the previous example, where $\partial W = H_C^+(S)$, it illustrates the full scope of (10.92).

The proof of case 2 is now virtually the same as for case 1: take $x \in H_C^+(S)$, seen as a specific component of the boundary of W. Since $W = I^+(H_C^+(S))$ there is a sequence (x_n) in $I^+(H_C^+(S))$ converging to x, for each x_n there is $y_n \in H_C^+(S)$ with $x_n \in I^+(y_n)$, and hence there are pd timelike curves γ_n from x_n to y_n, whose limit curve is the desired lightlike geodesic in $H_C^+(S)$. Since it is enough for Corollary 10.17.2, we assume $\mathrm{edge}(S) = \emptyset$, in which case the above construction can be repeated, so that this geodesic has no past endpoint in $H_C^+(S)$. □

Corollary 10.18 *If both the null curvature condition (see Theorem 6.15) and weak cosmic censorship hold (in that $I^-(\mathscr{I}^+)$ is globally hyperbolic, cf. Definition 10.8 and Theorem 10.9), then any future trapped surface S must lie entirely* within *the black hole region B^+.*

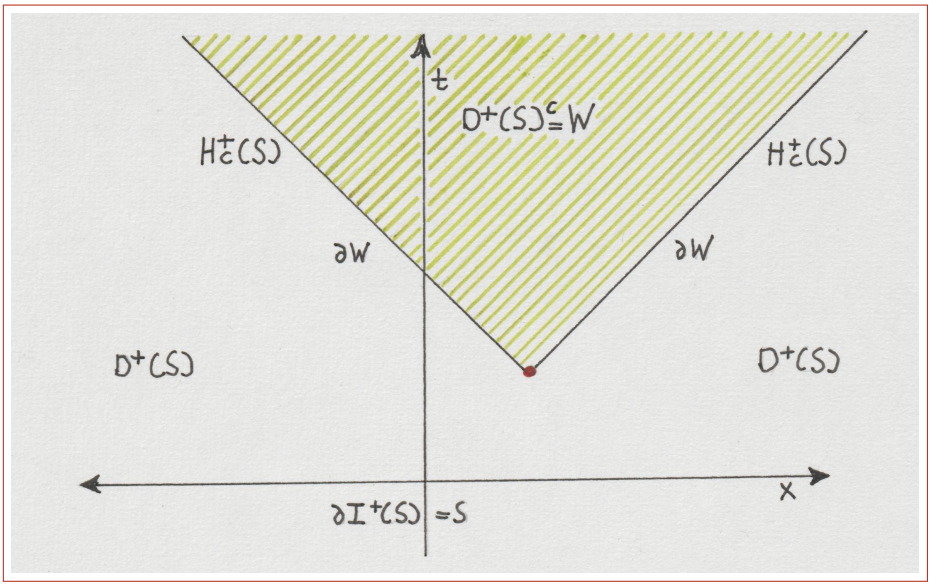
We just sketch the proof (by contradiction).[565] If S were to (partly) lie in $I^-(\mathscr{I}^+)$, then also part of $\partial I^+(S)$ lies in $I^-(\mathscr{I}^+)$ and hence some of the lightlike geodesics γ in Proposition 10.16.1, with past endpoint on S, would reach $\mathscr{I}^+$ and hence have infinite length. But the definition of a trapped surface excludes this, as in the proof of Penrose's singularity theorem (see §6.4). □

[562] Penrose (1972), p. 24. Other proofs are in Wald (1984), Theorem 8.1.6 and Galloway (2014), Proposition 3.4.

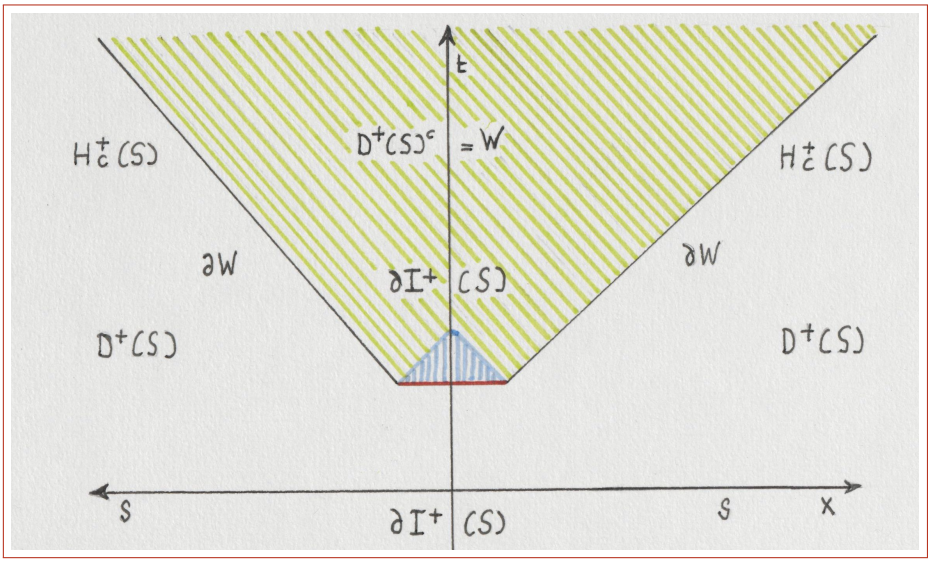
[563] Lemma 5.26 assumes that (M,g) is globally hyperbolic, but this assumption is not necessary here: the second bullet point in Lemma 5.40, which is case (ii) in Theorem 2.53 in Minguzzi (2019), can be excluded because we now have the sharper assumption $\gamma_n(b_n) = y$ for all n (as opposed to $\gamma_n(b_n) \to y$ in Lemma 5.26), so that Proposition 5.21 makes b finite. Alternatively, one can use Chruściel's limit curve lemma mentioned in footnote 223. Our argument is supposed to be a rigorous version of the corresponding proof in Geroch & Horowitz (1979), p. 234.

[564] The argument is due to Penrose (1972), proof of Theorem 5.12, pp. 44–45. For a different, very detailed proof see Minguzzi (2019), Theorem 3.24, compared to which the argument we give should be seen as heuristic.

[565] See Proposition 9.2.1 in Hawking & Ellis (1973), originally by Hawking (1972), corrected by Claudel (2000).



2d Minkowski space-time with $(1,1)$ *deleted, and the x-axis as a wannabe Cauchy surface S. Then* $D^+(S)$ *is the region between the x-axis and the two 45° lines, so that W is the shaded green area above these lines, which is excluded from* $D^+(S)$ *because any causal curve in W can disappear into the "singularity"* $(1,1)$ *instead of reaching S. Furthermore,* $I^+(S)$ *is the upper half plane without* $(1,1)$*. Thus* $\partial I^+(S) = S$*, and* $\partial W = H_C^+(S)$ *consists of the two 45 ° lines emanating from the deleted point* $(1,1)$*.*



Quinten space-time, where the closed red line segment from $(t,x) = (2,-1)$ *to* $(2,1)$ *is deleted from 2d Minkowski space-time, with wannabe Cauchy surface S again taken to be the x-axis. Then* $I^+(S)$ *is the upper half plane minus the dashed blue triangle with vertices* $(2,-1)$*,* $(2,1)$*, and* $(3,0)$*, including its interior, and* $D^+(S)\backslash S$ *is the open region enclosed between the x-axis and the two* $45°$ *lines connected by the red line (not included in* $D^+(S)$*). Furthermore,* $H_C^+(S)$ *consists of these* $45°$ *lines. Next,* $\partial I^+(S)\backslash S$ *consists of the blue upper sides of the triangle, and finally W is the region above the zig-zag pattern formed by* $H_C^+(S)$ *and* $\partial I^+(S)$*, with boundary* ∂W *as in* (10.92)*. Figures by Edith de Jong-de Liefste.*

10.8 Killing horizons and surface gravity

We turn to the third type of black hole horizon, which is important for the development of black hole thermodynamics. Unlike the previous two it is only defined if the metric is stationary.[566]

Definition 10.19 *A* **Killing horizon** *in a space-time (M,g) is a connected null hypersurface $H_K \subset M$ with a normal vector field N (which by definition is lightlike on H_K) that can be extended to a Killing vector field X on some nbhd of H_K, in which nbhd it is lightlike solely on H_K. Equivalently, a Killing horizon for a Killing vector field X defined on some open subset $U \subset M$ is a connected hypersurface $H_K \subset U$ that coincides with a connected component of the subset of U where X is lightlike (and hence nonzero), and X is normal (and hence tangent) to H_K.*

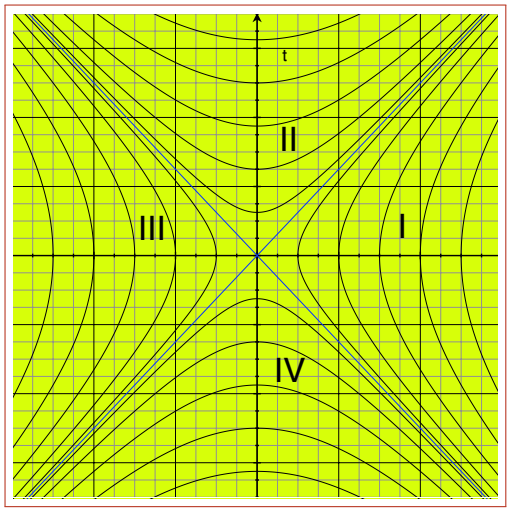
For example, in Schwarzschild black holes $X = \partial_t$ is timelike outside the hole, lightlike on the event horizon, making this also a Killing horizon, and spacelike after crossing the horizon inwards (see below for Kerr and Reissner–Nordström). This situation is surprisingly well mimicked by

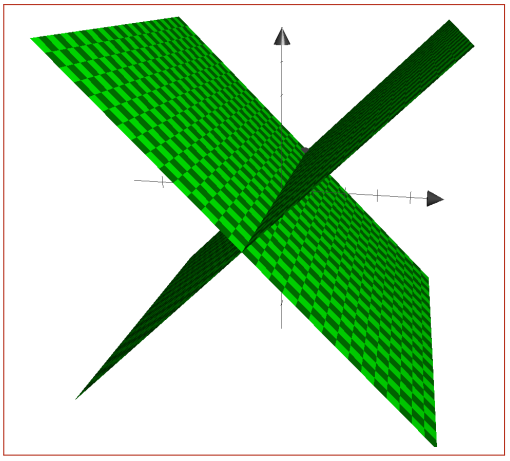
$$X = x\partial_t + t\partial_x \tag{10.93}$$

in $2d$ Minkowski space-time (or indeed in any dimension). This is a boost generator and hence an isometry,[567] whose flow is well known from special relativity and is given by

$$t(s) = t_0 \cosh s + x_0 \sinh s; \qquad x(s) = x_0 \cosh s + t_0 \sinh s, \tag{10.94}$$

where $s \in \mathbb{R}$ (i.e. X is complete).[568] Some of the flow line are displayed in the left-hand figure.





Left: *Flow lines (black) and bifurcate Killing horizon (blue) of the vector field $X = x\partial_t + t\partial_x$ in $2d$ Minkowski space. Clearly, X is timelike in the regions I and III, spacelike in regions II and IV, and lightlike on the horizon, just like ∂_t in the Kruskal case. The bifurcation surface is the origin.*
Right: *Bifurcate Killing horizon of the same vector field X in $3d$. The bifurcation surface is the y-axis (pointing out of the page). The bifurcation surface always has codimension 2 and e.g. for Kruskal is S^2.*

[566] For more information see e.g. Chruściel (2020), §4.3, Aretakis (2013), §5.6, and Poisson (2004), chapter 5.
[567] Killing's equation $X_{\mu;\nu} + X_{\nu;\mu} = 0$ reads $X_{\mu,\nu} + X_{\nu,\mu} = 0$, with $X^0 = x$, $X^1 = t$ and hence $X_0 = -x$, $X_1 = t$.
[568] This flow is not parametrized by proper time τ. In region I, the ***Rindler wedge***, putting $t_0 = 0$ this is achieved by $t(\tau) = x_0 \sinh(\tau/x_0)$ and $x(\tau) = x_0 \cosh(\tau/x_0)$. This gives the well-known constant acceleration $1/x_0^2$. See e.g. Misner, Thorne, & Wheeler, §6.6. In this context the $x \geq 0$ part of the horizon is called the ***Rindler horizon***: it represents the boundary of what can be (causally) known by the accelerating observers in region I (Rindler, 1956).

Since $\eta(X,X) = t^2 - x^2$, the Killing horizons of X are the following lines (or hypersurfaces):

$$x = t, x > 0; \qquad x = t, x < 0; \qquad x = -t, x > 0; \qquad x = -t, x < 0. \tag{10.95}$$

These combine into a cross $|x| = |t|$ (*without the origin in 2d and without the y-axis in 3d*); the four (open) regions I, II, III, IV enclosed by the sides of the cross resemble a Kruskal diagram.[569] This cross has an interesting structure, which again is shared e.g. by Kruskal space-time:

Definition 10.20 *A* **bifurcate Killing horizon** *in a space-time is a union of four (connected) Killing horizons (for the same Killing vector field X) connected by a submanifold $\mathscr{S}$ of dimension two (generally of* codimension two*) on which X vanishes, called the* **bifurcation surface***, and from which each of the four horizons emanates in a lightlike direction orthogonal to $\mathscr{S}$.*

This implies that a bifurcation surface $\mathscr{S} \subset M$ is spacelike. Conversely, $\mathscr{S}$ determines a bifurcate Killing horizon, as follows.[570] Suppose a Killing vector field X vanishes precisely on a two-dimensional spacelike submanifold $\mathscr{S}$. By Minkowski geometry (cf. the end of §4.6 and §6.3), at each $x \in \mathscr{S}$ the tangent space T_xM has a basis $(L, \underline{L}, e_1, e_2)$, where L and $\underline{L}$ are lightlike, preferably normalized as in (6.58), (e_1, e_2) is a basis of $T_x\mathscr{S} \subset T_xM$ (so that e_1 and e_2 are spacelike), and L and $\underline{L}$ are orthogonal to e_1 and e_2 (and hence to $T_x\mathscr{S}$). Since $X = 0$ on $\mathscr{S}$, its flow ψ_t leaves $\mathscr{S}$ pointwise invariant, so that its pushforward $(\psi_t)_* \equiv T\psi_t$ maps each tangent space T_xM into itself ($x \in \mathscr{S}$). As X is a Killing vector field, each ψ_t is an isometry of (M,g), and each $T_x\psi_t$ is an isometry of T_xM. In particular, $T_x\psi_t(L)$ must be lightlike, so that it must be proportional to either L or $\underline{L}$. Since $T_x\psi_0 = \mathrm{id}$ and hence $T_x\psi_0(L) = L$, proportionality to $\underline{L}$ is impossible by continuity in t. Hence there must be some function f such that

$$T_x\psi_t(L) = f(t)L. \tag{10.96}$$

Consider the geodesic $\gamma_L^{(x)}$ for which $\gamma_L^{(x)}(0) = x$ and $\dot{\gamma}_L^{(x)}(0) = L$. In general,

$$\gamma_{\psi_* Y}^{(x)}(\tau) = \psi(\gamma_Y^{(x)}(\tau)); \qquad \gamma_{sY}^{(x)}(\tau) = \gamma_Y^{(x)}(s\tau), \tag{10.97}$$

where Y is any element of T_xM and ψ is any isometry. The equation on the left follows because both sides are geodesics (this requires ψ to be an isometry, since arbitrary diffeomorphisms would not map geodesics to geodesics) with the same initial point $\psi(x)$ and tangent vector $\psi_* Y$ at that point. Taking $Y = L$ and $\psi = \psi_t$, the flow of X, then shows that, for the above $f(t)$,

$$\psi_t(\gamma_L^{(x)}(\tau)) = \gamma_L^{(x)}(f(t)\tau), \tag{10.98}$$

so that ψ_t maps $\gamma_L^{(x)}$ to itself. This is only possible if X is proportional to $\dot{\gamma}_L$ throughout $\gamma_L^{(x)}$, which in turn implies that X is lightlike throughout $\gamma_L^{(x)}$. Defining $H_K^+ := C$ as in (6.61), that is, as the union of all fd lightlike geodesics emanating from $\mathscr{S}$ with tangent L (assuming that the above basis is defined smoothly all over $\mathscr{S}$), we obtain a Killing horizon. The same construction works with $-L$ and with $\pm\underline{L}$, yielding four Killing horizons $H_K^\pm$ and $\underline{H}_K^\pm$, which combine with $\mathscr{S}$ to form a bifurcate Killing horizon. Note that by the same arguments the space between these horizons is filled with geodesics whose tangents, still proportional to X, cannot be lightlike.

Conversely, we will show that if the surface gravity κ on some Killing horizon H_K, which we will now introduce, is strictly nonzero, then H_K extends to a bifurcate Killing horizon.

[569] There is a different way of looking at these Killing horizons, which has an analogue for black holes: if γ is any of the curves in region I, then the $x = t$ line equals $\partial I^-(\gamma)$. Similarly, $x = -t$ equals $\partial I^-(\gamma)$ for any γ in region III.

[570] What follows explicates an argument in Chruściel (2020), §4.3.2.

Proposition 10.21 *Let H_K be a Killing horizon for some Killing vector field X. Then* on H_K,

$$\nabla_X X = \kappa X, \tag{10.99}$$

for some function κ defined on H_K, called the **surface gravity** *of the horizon H_K. It satisfies*

$$X\kappa = 0. \tag{10.100}$$

Eq. (10.100) means that κ is constant along the null generators of H_K (i.e. the lightlike pre-geodesics with tangent X); in §10.12 we give conditions under which κ is even constant on H_K (which is the *zeroth law of black hole thermodynamics*). Note that X is orthogonal but also tangent to the null hypersurface H_K (see §4.6), so that $X\kappa$ is well defined even if κ is not defined outside H_K. Clearly, the flow of X on H_K is geodesic iff $\kappa = 0$, in which case H_K is called ***degenerate***. Likewise, a Killing horizon is ***non-degenerate*** if κ is nonzero throughout H_K.

Proof. Since $g(X,X) = 0$ *on* H_K one has $Zg(X,X) = 0$ *for each Z tangent to* H_K. Therefore,

$$Zg(X,X) = (\nabla_Z g)(X,X) + g(\nabla_Z X, X) + g(X, \nabla_Z X) = 2g(\nabla_Z X, X) = 0. \tag{10.101}$$

Taking $Y = X$ in (9.122) and using (10.101) gives $g(\nabla_X X, Z) = 0$ for each Z tangent to H_K, which implies that $\nabla_X X$ must be normal to H_K and hence proportional to X. This proves (10.99).

We derive (10.100) from an identity for any Killing vector field X and $Y, Z \in \mathfrak{X}(M)$, viz.[571]

$$\nabla_Y \nabla_Z X - \nabla_{\nabla_Y Z} X = \Omega(Y,X)Z := ([\nabla_Y, \nabla_X] - \nabla_{[Y,X]})Z. \tag{10.102}$$

Putting $W = \nabla_X X$ in torsion-freeness $\mathscr{L}_X W = \nabla_X W - \nabla_W X$, and $Y = Z = X$ in (10.102), gives

$$\mathscr{L}_X(\nabla_X X) = 0. \tag{10.103}$$

Using (10.99), this gives $\mathscr{L}_X(\kappa X) = 0$, i.e. $(X\kappa)X + \kappa \mathscr{L}_X X = (X\kappa)X = 0$, whence (10.100).

We also give an equivalent but self-contained proof in coordinates, starting from the identity

$$\nabla_\mu \nabla_\nu X^\alpha = R^\alpha_{\nu\mu\beta} X^\beta, \tag{10.104}$$

valid for any Killing vector field X. Using (9.122), i.e. $\nabla_\mu X_\nu + \nabla_\nu X_\mu = 0$, and (4.13) gives

$$\begin{aligned}
\nabla_\mu \nabla_\nu X_\alpha &= -\nabla_\mu \nabla_\alpha X_\nu = -\nabla_\alpha \nabla_\mu X_\nu + R_{\nu\beta\alpha\mu} X^\beta = \nabla_\alpha \nabla_\nu X_\mu + R_{\nu\beta\alpha\mu} X^\beta \\
&= \nabla_\nu \nabla_\alpha X_\mu + R_{\mu\beta\alpha\nu} X^\beta + R_{\nu\beta\alpha\mu} X^\beta = -\nabla_\nu \nabla_\mu X_\alpha + R_{\mu\beta\alpha\nu} X^\beta + R_{\nu\beta\alpha\mu} X^\beta \\
&= -\nabla_\mu \nabla_\nu X_\alpha + R_{\alpha\beta\mu\nu} X^\beta + R_{\mu\beta\alpha\nu} X^\beta + R_{\nu\beta\alpha\mu} X^\beta.
\end{aligned} \tag{10.105}$$

From (4.24) and (4.36) - (4.38) we then obtain (10.104). Furthermore, also for later, we have

$$\kappa^2 = -\tfrac{1}{2} \nabla^\nu X^\alpha \nabla_\nu X_\alpha, \tag{10.106}$$

valid on the Killing horizon H_K. To derives this, use (8.94), which in coordinates reads

$$(\nabla_\nu X_\mu) X_\rho + (\nabla_\mu X_\rho) X_\nu + (\nabla_\rho X_\nu) X_\mu = 0, \tag{10.107}$$

contract with $\nabla_\nu X^\mu$, and use (10.99) and (3.74). Applying $X^\alpha \nabla_\alpha$ to (10.106), eq. (10.104) gives

$$2\kappa X^\alpha \partial_\mu \kappa = -R^\alpha_{\nu\mu\beta} X^\mu X^\beta \cdot \nabla^\nu X_\alpha = 0. \qquad \square$$

[571] For a proof see e.g. Aretakis (2013), page 87. Our subsequent coordinate proof follows Poisson (2004), §5.5.2.

As explained after (3.48), where we restrict the setting to H_K (which can be done since X is tangent to it), eq. (10.99) shows that the flow of X can be reparametrized to make it *geodesic*. Since X is lightlike on H_K by definition, the ensuing flow consists of lightlike geodesics. Hence *the lightlike geodesics ruling the null hypersurface H_K according to Proposition 6.11 are reparametrized flow lines of X*. Suppose $X = fL$ for some function f defined at least on H_K, and L a null vector field on H_K so that $\nabla_L L = 0$, i.e. its flow is geodesic. Then $\kappa = Lf$, i.e.

$$f(\tau) = \kappa\tau + c; \qquad \Rightarrow \qquad X(\tau) = (\kappa\tau + c)L, \tag{10.108}$$

along the geodesic flow $\tau \mapsto \gamma_L(\tau)$ of L, where τ is an affine parameter. If $\kappa \neq 0$, then X vanishes at $\tau = -c/\kappa$, which means that the Killing horizon has hit the bifurcation surface of a bifurcate Killing horizon, provided that γ can be extended that far. We close with some examples.[572]

- The surface gravity for the Killing vector field (10.93) in Minkowski space-time is given by $\kappa = \pm 1$ on the $x = \pm t$ components of the Killing horizon.

- In the *Schwarzschild solution* (9.15) in the original coordinates (t, r, θ, φ) the obvious Killing vector field is $X = \partial_t$, but since these coordinates do not apply exactly where things become interesting, namely at $r = 2m$, we switch to ingoing Eddington–Finkelstein coordinates (v, r, θ, φ), with metric (9.44), and take (or: write) $X = \partial_v$, which is the same vector field (as a computation shows). The metric (9.44) shows that X is timelike for $r > 2m$, lightlike at $r = 2m$, and spacelike throughout $0 < r < 2m$. In particular, the future event horizon H_E^+ defined in Theorem 9.1 is a Killing horizon, too. We may then compute κ from its definition (10.99), which simply comes down to $\kappa = \Gamma^v_{vv}$. This can be computed from (4.15) and (9.44), yielding $\Gamma^v_{vv} = m/r^2$, which at $r = 2m$ gives $\kappa = 1/4m$.

- In the *Kruskal solution* (9.55), for the Killing vector field (9.63) one finds that $\kappa = 1/4m$ on the two SW–NE Killing horizons (including the future or black hole event horizon just treated), and $\kappa = -1/4m$ on the SE–NW ones (and hence in particular on the past or white hole event horizon). The bifurcation surface is the two-sphere at the origin.

- The *Reissner–Nordström metric* (9.88) with $0 < |e| \leq m$ and $X = \partial_t = \partial_v$ has two Killing horizons, which coincide with the inner and outer horizons $H_\pm$ of Theorem 9.2. This follows from (9.94), which makes ∂_v lightlike iff $h(r) = 0$, which is the case at $r = r_\pm$. The surface gravities $\kappa_\pm$ coincide with those already labeled as such in (9.92) and (9.106). In the extremal case $|e| = m > 0$ the surface gravity on the single remaining Killing horizon = event horizon = Cauchy horizon vanishes. For $|e| > m > 0$ there is no horizon at all.

- The *Kerr metric* (9.110) has a second Killing vector ∂_φ, apart from ∂_t, which again coincides with ∂_v as used in (9.137). This additional symmetry makes the choice of X ambiguous, but the Killing horizon of X coincide with the horizon H_+ in Theorem 9.3 if

  $$X_+ := \partial_v + \Omega_+ \partial_\varphi, \tag{10.109}$$

  see (9.143). With this choice of X, the surface gravity at $H_+ = H_E^+$ is given by

  $$\kappa_+ = \tfrac{1}{2}\frac{r_+ - r_-}{r_+^2 + a^2} = \tfrac{1}{2}\frac{\sqrt{m^2 - a^2}}{m^2 + m\sqrt{m^2 - a^2}}, \tag{10.110}$$

  at least if $0 < |a| \leq m$. The extremal case $|a| = m > 0$ has $\kappa = 0$ (and *vice versa*), and in the ultrafast case $|a| > m > 0$ there is no horizon whatsoever, but a naked singularity.

[572] The coincidence of Killing horizons and event horizons is no coincidence and will be taken up in §10.10.

10.9 Black hole uniqueness theorems: Static case

Uniqueness theorems in GR, more specifically in the theory of black holes, typically refer to claims to the effect that under certain assumptions appropriate to the black hole setting, at least the space-time outside the event horizon must be (locally) isometric to one of the classical exact solutions, such as Schwarzschild, Reissner-Nordström, Kerr, or Kerr–Newman. Thus the uniqueness theorems formalize what (following Wheeler) is often called the "no hair" property:

> Perhaps the greatest surprise from the golden age [i.e. 1963–1975] was general relativity's insistence that all properties of a black hole are precisely predictable from just three numbers: the hole's mass, its rate of spin, and its electric charge.[573] From those three numbers, if one is sufficiently clever at mathematics, one should be able to compute, for example, the shape of the hole's horizon, the strength of its gravitational pull, the details of the swirl of space-time around it, and its frequencies of pulsation. (Thorne, 1994, p. 327)

The scope of the black hole uniqueness theorems ranges from the early (misnamed) ***Birkhoff theorem*** from 1923, see below, to Penrose's all-encompassing ***final state conjecture***:[574]

> A body, or collection of bodies, collapses down to a size comparable to its Schwarzschild radius, after which a trapped surface can be found in the region surrounding the matter. Some way outside the trapped surface region is a surface which will ultimately be the absolute event horizon. But at present, this surface is still expanding somewhat. Its exact location is a complicated affair and it depends on how much more matter (or radiation) ultimately falls in. We assume only a finite amount falls in and that GIC is true. Then the expansion of the absolute event horizon gradually slows down to stationarity. Ultimately the field settles down to becoming a Kerr solution (in the vacuum case) or a Kerr–Newman solution (if a nonzero net charge is trapped in the "black hole"). (Penrose, 1969, pp. 1157–1158)

Here GIC refers to what Penrose (1969) called the *Generalized Israel Conjecture*, i.e.,

> if an absolute event horizon develops in an asymptotically flat space-time, then the solution exterior to this horizon approaches a Kerr–Newman solution asymptotically with time. (Penrose, 1969, pp. 1156)

In the static case, which Israel himself proved (albeit under very restrictive assumptions) in two papers that launched the modern era,[575] this means hat the solution exterior to this horizon *equals* the Reissner-Nordström solution (and hence the Schwarzschild solution in the vacuum case). This requires the inference of spherical symmetry from staticity, which is a (much more difficult) converse of the inference of staticity from spherical symmetry in Birkhoff's theorem.

[573] The latter seems zero in astrophysical reality (where nonetheless black holes in active galactic nuclei are surrounded by magnetic fields), unless 't Hooft's idea that elementary particles are tiny black holes is viable.

[574] 'The conjecture is extremely open, in the sense that even a reasonable formulation is unknown.' (Wong, 2009)

[575] These are Israel (1967, 1968). Overviews of the uniqueness theorem, including references to the original literature (some of which will also be cited below) include Hawking & Ellis (1973), §9.3, Carter (1986), Heusler (1996), and Chruściel, Lopes Costa, & Heusler (2012). The history of the theorems is discussed in first-hand accounts by Israel (1987), Carter (2006), Thorne (1994), chapter 7, and Robinson (2009). Briefly, the "no hair" conjecture originated in the Moscow from work by Ginzburg on quasars and independently Doroshkevich, Novikov, and Zeldovich on deformations of black holes. In 1965 Novikov presented this work at the GR4 conference in London, through which it reached the West, where the idea was picked up by Wheeler and his former students like Thorne and Misner, by Israel, and subsequently, via the latter, by Carter, Hawking, and others.

Since a complete treatment of the uniqueness theorems would require an entire monograph, our aim here is just to generate some feeling for these theorems by discussing a few special cases in some detail, and even those, for clarity, under stronger assumptions than strictly needed, and with mere outlines of the main steps in the proofs (which would take pages *per step* if done in detail). Stronger up-to-date results will be mentioned along the way without proof.

As already mentioned, the first uniqueness theorem for black holes was Birkhoff's:[576]

Theorem 10.22 *Any spherically symmetric solution to the vacuum Einstein equations is locally isometric to the Schwarzschild solution (i.e. for all values of $r > 0$).*

The most remarkable aspect of this theorem is that spherical symmetry *implies* staticity, but even if staticity is *assumed* the conclusion would be non-trivial. Of course, everything is predicated on the exact definition of spherical symmetry. Using coordinates (x^μ) where the rotation group $SO(3)$ acts trivially on $x^0 = t$, and acts in the usual way on (x^1, x^2, x^3), $SO(3)$-invariance forces

$$g_{ij} = A\delta_{ij} + Bx^i x^j; \qquad g_{0i} = Cx^i; \qquad g_{00} = -D, \tag{10.111}$$

where A, B, C, and D depend on $x_1^2 + x_2^2 + x_3^2$ and $x^0 = t$ only. Replacing (x^1, x^2, x^3) by spherical coordinates (r, θ, φ) but redefining r so that the volume element of S_r^2 is $r^2 d\Omega$, cf. (9.17), gives

$$g = -Edt^2 + 2F\, dr\, dt + G^2 dr^2 + Hd\Omega, \tag{10.112}$$

where E, F, G, and H depend on r and t only. A further coordinate transformation then yields

$$g = I(u,t)du^2 + 2J(u,r)\, dr\, dt + K(u,r)d\Omega, \tag{10.113}$$

which the vacuum Einstein equations then force into the Schwarzschild metric (9.44) or (9.45) in Eddington–Finkelstein coordinates (the lengthy and dull computations are left to the reader).

But the above concept of spherical symmetry was coordinate-dependent. We can do better:

Definition 10.23 *A space-time (M, g) is* **spherically symmetric** *if $SO(3)$ is a nontrivial subgroup of the group $\mathrm{Iso}(M, g)$ of its isometries and the orbits of $SO(3)$ are isometric to spacelike two-spheres S_r^2 of some radius $r > 0$ (endowed with the usual round metric).*[577]

The following–very technical–lemma will do much of the work. Part 1 may sound trivial given Definition 10.23, but its thrust lies in the precise meaning of a foliation (see footnote 578).

[576] What is called ***Birkhoff's theorem*** was also independently discovered by Jebsen (1921), Alexandrow (1923), and Eisland (1925). See Johansen & Ravndal (2006) and Ehlers & Krasiński (2006). The name-giving source is Birkhoff (1923), which is sometimes cited as Birkhoff & Langer (1923); the cover says 'By George David Birkhoff, PhD, with the coorperation of Rudolph Ernest Langer, PhD'. Most GR textbooks contain computations supporting the theorem, e.g. Misner, Thorne, & Wheeler (1973), §23.2 and §32.2, is very clear. The *Ansatz* (10.111) is taken from Deser & Franklin (2005) and the subsequent (original) use of Eddington–Finkelstein coordinates is due to van Oosterhout (2019), which contains a detailed derivation of (9.44) or (9.45) from (10.113); this has the advantage of not being limited to $r > 2m$. Complete and rigorous geometric proofs are hard to find. We follow Hawking & Ellis (1973), Appendix B, which relies on Lemma 10.24 due to Schmidt (1967). See also van Oosterhout (2019). Birkhoff's theorem was extended to electrovac space-times by Hoffmann (1932ab).

[577] This excludes Minkowski space-time $(\mathbb{M}, \eta)$, which near $r = 0$, is not foliated by two-spheres! Birkhoff's theorem produces $(\mathbb{M}, \eta)$ without the t-axis, which can be added by moving back to Cartesian coordinates.

Lemma 10.24 *1. A spherically symmetric space-time (M,g) is foliated by two-spheres.*[578]

2. Each $x \in M$ has a nbhd U_x containing a 2d submanifold $\mathscr{N}_x$ through x that intersects each orbit (i.e. two-sphere S^2) overlapping with U_x exactly once and does so orthogonally.

3. For any two (nearby) orbits $\mathscr{O}$ and $\mathscr{O}'$ the map $\mathscr{O} \to \mathscr{O}'$ that sends $x \in \mathscr{O}$ to $\mathscr{N}_x \cap \mathscr{O}'$ (provided this is nonemtpy, in which case it has one element) is a conformal diffeomorphism whose conformal factor Ω is constant on $\mathscr{O}$ (i.e. depends on $\mathscr{O}$ and $\mathscr{O}'$ alone).

Visualizable examples in $n = 3$ include $\mathbb{R}^3\backslash\{0\}$, seen as Euclidean space (minus the origin) foliated by two-spheres ($G = SO(3)$), and $\mathbb{R}^3\backslash\{x^1 = x^2 = 0\}$, seen as Minkowski space-time $\mathbb{M}_3$ in $d = 2+1$, foliated by circles in planes with $x^0 =$ constant ($G = SO(2)$). In the first example of the previous footnote $\mathscr{N}_x$ is (locally) simply the radial line through x. In the second, it is (locally) the plane defined as the product of the radial line through x and the t-axis.

Proof. If a Lie group G acts smoothly on M and $\dim(G_x)$ is constant, where

$$G_x = \{g \in G \mid gx = x\}, \tag{10.114}$$

is the stabilizer of x, then the associated vector fields defined by (8.246) define a foliation of M, whose leaves are the connected components of the G-orbits $\mathscr{O}_x = Gx$. This is the situation here, with $G = SO(3)$ and $G_x \cong SO(2)$, and connected orbits $\cong S^2$. This proves the first claim.

For the second claim,[579] the more general fact is that there is such an $\mathscr{N}_x$ provided each $\psi \in G_x$ (different from the identity) satisfies $\psi_* X = X$ iff $X = 0$, for $X \in T_x\mathscr{O}$. This assumption certainly holds if the $SO(3)$ orbits are all two-spheres, in which case $G_x \cong SO(2)$ rotates $T_xS^2 \cong \mathbb{R}^2$ (note that since $\psi(x) = x$, the pushforward ψ_* maps T_xM to itself). To prove this, define $\mathscr{N}_x$ as the submanifold generated by all geodesics emanating from x with tangents orthogonal to the orbit $\mathscr{O}_x$. The slice theorem for compact Lie group actions gives the required nbhd U_x (where $S = \mathscr{N}_x \cap U_x$ acts as the slice). We now show that if $X \perp T_x\mathscr{O}_x$ and $\psi \in G_x$, then $\psi_* X = X$. Indeed, if $Y = \psi_* \neq X$, then (because ψ is an isometry) $Y \perp T_x\mathscr{O}_x$, and the different geodesics $\gamma_X^{(x)}$ and $\gamma_Y^{(x)}$ both intersect some orbit $\mathscr{O}$ near x. But since ψ is an isometry,

$$\gamma_Y^{(x)}(t) = \gamma_{\psi_* X}^{(x)}(t) = \psi(\gamma_X^{(x)}(t)), \tag{10.115}$$

and so $\mathscr{O}$ would intersect $\mathscr{N}_x$ in more than one point, contradicting the slice theorem. Now take $y = \gamma_X^{(x)}(s) \in U_x$ for some $s \neq 0$. By the same calculation, $\psi(y) = y$, so $G_x \subseteq G_y$ and hence $\psi_*(Y) = Y$ for each $Y \perp T_y\mathscr{O}_y$ (where this time, $\psi_* : T_yM \to T_yM$). Now decompose

$$\dot{\gamma}_X^{(x)}(s) = Y_1 + Y_2, \tag{10.116}$$

with $Y_1 \perp T_y\mathscr{O}_y$ and $Y_2 \in T_y\mathscr{O}_y$. We know that $\psi_*(Y_1) = Y_1$, and $\psi_*(Y_2) \neq Y_2$ would lead to a similar contradiction with the slice theorem as previously at x, so $\psi_*(Y_2) = Y_2$ and hence $Y_2 = 0$.

[578] A k-dimensional foliation of an n-dimensional manifold M, where $0 < k < n$, may be defined as a subbundle $E \subset TM$ of rank k that is *involutive* in the sense that if X, Y are sections of E, then so is their Lie bracket $[X,Y]$. In that case, M is the disjoint union of the *leaves* of the foliation, which are (immersed) connected submanifolds $\mathscr{L}$ of M such that $T_x\mathscr{L}_x = E_x$ at each $x \in M$ (where $\mathscr{L}_x$ is the leaf through x). The nontrivial fact we need is that, for a general foliation, each $x \in M$ has a nbhd U and associated chart (U,φ) with ensuing coordinates (x^i) such that $\mathscr{L}_x \cap U$ is given by $x^1 =$ constant, $\ldots$, $x^{n-k} =$ constant. See e.g. Guillemin & Sternberg (1984), §27.

[579] The general case is due to Schmidt (1967), Theorem 1. For the slice theorem used in the proof below see e.g. Guillemin & Sternberg, 1984, Proposition 27.2.

Thus $\gamma_X^{(x)}$ and hence $\mathscr{N}_x$ intersects all orbits in U_x orthogonally, [580] proving the second claim.

Call the map in the third claim $f : \mathscr{O} \to \mathscr{O}'$; since the orbits are compact, f can indeed be defined on all of $\mathscr{O}$. This map is well defined by the previous claim and it is a local diffeomorphism by the properties of the exponential map. For any $\varphi \in G$ one has $f \circ \varphi = \varphi \circ f$ by computations on geodesics as in the previous step. Now choose an orthonormal basis $(e_1, \dots, e_k)$, with $k = 2$ in the case $G = SO(3)$ and $G_x = SO(2)$ at hand, obtained from a unit vector $u \in T_x\mathscr{O}$ by $e_i = \psi_i' u$ for suitable $\psi_i \in G_x$. Then all e_i are unit vectors, and by G-equivariance (as just noted) these are mapped into an orthogonal basis $u_i = \psi_i(f_* e)$, which also consists of vectors of the same length. Linear conformal transformations are compositions of rotations, reflections, and dilations,[581] and hence f is conformal. Since G consists of isometries and acts transitively on each orbit, a simple computation shows that the conformal factor is constant on $\mathscr{O}$. □

By foliation theory, it is now possible to introduce coordinates (t, r, θ, φ) on U_x such that:

- each orbit $\mathscr{O}$ is given by $t =$ constant and $r =$ constant;
- each normal surface $\mathscr{N}_x$ is given by $\theta =$ constant and $\varphi =$ constant.

Here (θ, φ) are spherical coordinates on S^2. This yields (10.112), with which we close our discussion of Birkhoff's theorem; as already mentioned, we will not give the explicit computations that lead from (10.112) to the Schwarzschild metric (but see §9.2, which assumed staticity).

Spherical symmetry is a very strong assumption, and so it is interesting to know that the Schwarzschild solution can also be inferred from a very different set of assumptions, which now *includes* staticity. The conclusion of the theorem is both global and restricted to the exterior region $r > 2m$, which requires the use of a manifold with boundary in the assumptions.[582]

Theorem 10.25 *Let (M, g) be a one-ended static asymptotically flat space-time (cf. Definition 8.4) with metric (8.96), such that $L > 0$ in $\mathrm{int}(M)$ and $L = 0$ on ∂M, where $M = \mathbb{R} \times \Sigma$ and $\partial M = \mathbb{R} \times \partial\Sigma$, with $\partial\Sigma$ compact. If the metric solves the vacuum Einstein equations (8.100) - (8.101) and (M, g) is maximally extended up to its boundary, then (M, g) is isometric to the exterior region $r \geq 2m$ of Schwarzschild space-time (9.46) with metric (9.15) having $m > 0$. In particular, the boundary ∂M is connected and coincides with the future event horizon H_E^+.*

Despite its strong assumptions, this theorem is quite remarkable, since it not only shows that the Schwarzschild metric is the only static vacuum black hole space-time (which by definition is asymptotically flat) with smooth event horizon, but it also shows that multiple black hole configurations (which would form a space-time with disconnected boundary and hence are excluded by the theorem) in vacuum cannot be static. In order to understand its assumptions, it is worth recalling that $L^2 = -g(\partial_t, \partial_t)$, where ∂_t is the (usual) timelike Killing vector field of a static space-time, so that the vanishing of L at ∂M makes the latter a Killing horizon, which *a posteriori* is identified with the event horizon of a Schwarzschild black hole, cf. Theorem 9.1.

---

[580] Since we now know that $\dot\gamma_X^{(x)}(s) \perp T_y\mathscr{O}_y$ we may run the geodesic (and the argument) the other way round, obtaining the inclusion $G_y \subseteq G_x$, and hence $G_y = G_x$. Since $\psi_*(X) = X$ for $\psi \in G_x$ and $X \perp T_x\mathscr{O}_x$, we also have $\psi(\gamma_X^{(x)}(t)) = \gamma_X^{(x)}(t)$, so that $\mathscr{N}_x$ is *pointwise* invariant under G_x (as the examples indeed illustrate).

[581] This is ***Liouville's theorem***, see e.g. Akivis & Goldberg, 1996, Theorem 1.1.1.

[582] This is a version of ***Israel's theorem*** from 1967, see footnote 575, due to Bunting & Masood-ul-Alam (1987). Israel assumed the boundary ∂M to be connected and also made several other superfluous assumptions. See also Heusler (1996), §9.2, Schoen (2009), Lecture 11, and the references in footnote 575 for historical context.

The proof (which we only sketch) aims at recovering the *spatial* Schwarzschild metric g_S from certain characteristic properties, upon which (8.96) gives the space-time metric on $M'_S = \mathbb{R} \times \Sigma'_S$. The spatial metric g_S is defined on $\Sigma'_S := \mathbb{R}^3 \backslash ((0,2m] \times S^2)$ and is given by

$$\tilde{g}_S = L(r)^{-2} dr^2 + r^2 (d\theta^2 + \sin^2\theta d\varphi^2); \qquad L(r) = \sqrt{1 - 2m/r}, \tag{10.117}$$

Listed in the order of the chain of deduction, these characteristic properties of $\tilde{g}_S$ are as follows:

1. Beyond the generic properties $\tilde{g}_{ij} = \delta_{ij} + O(1/r)$ and $L = 1 + O(1/r)$ from Definition 8.4, the specific asymptotics of $g = g_S$, with $g_{00} = L^2$ and $\hat{g}_{ij} := L^2 \tilde{g}_{ij}$, also satisfy

$$L = 1 - \frac{m}{r} + O\left(\frac{1}{r^2}\right); \qquad \hat{g}_{ij} = \delta_{ij} + O\left(\frac{1}{r^2}\right); \qquad (r \to \infty), \tag{10.118}$$

   whereas Definition 8.4 only requires $\hat{g}_{ij} - \delta_{ij} + O(1/r)$. In general, for our asymptotically flat spatial metric $\tilde{g}$, the asymptotics expressed by eqs. (10.118) follow from the Einstein equations (8.102), roughly speaking as follows.[583] The second, $\tilde{\Delta} L = 0$, where $\tilde{\Delta}$ is the Laplacian defined by $\tilde{g}$, has as lead term $\Delta L = 0$, where $\Delta = \partial_x^2 + \partial_y^2 + \partial_z^2$ is the usual flat Laplacian. In $3d$ flat space Laplace's equation is solved by

$$L = C - m/r, \tag{10.119}$$

   where m and C are constants. Then $L \to 1$ forces $C = 1$ and the error term in passing from Δ to $\tilde{\Delta}$ gives the first entry in (10.118). Next, in terms of $U = \ln(L)$, eqs. (8.102) read

$$\hat{R}_{ij} = 2\partial_i U \partial_j U; \qquad \hat{\Delta} U = 0, \tag{10.120}$$

   where $\hat{R}_{ij}$ and $\hat{\Delta}$ are the Ricci tensor and the Laplacian defined by $\hat{g}$, respectively. The asymptotics of L give $U = O(1/r)$ and hence $\hat{R}_{ij} = O(1/r^4)$, as $r \to \infty$. In harmonic coordinates (where $\hat{\Delta} x^i = 0$, i.e. $\Gamma^i := \hat{g}^{jk} \hat{\Gamma}^i{}_{jk} = 0$), this yields the second part of (10.118).

2. *The spatial metric $\tilde{g}_S$ is conformally flat.*[584] This follows either from a reparametrization

$$r = \rho(1 + m/(2\rho))^2; \tag{10.121}$$

$$\tilde{g}_S = (1 + m/(2\rho))^4 (d\rho^2 + \rho^2 d\Omega), \tag{10.122}$$

   or by computing the Cotton tensor (4.121) for the metric (10.117), which gives zero.

   We now prove that our Riemannian manifold with boundary $(\Sigma, \tilde{g})$, as defined through the assumptions in Theorem 10.25 plus Definition 8.4, must be conformally flat. To this end, we first perform a useful but partly unsuccessful manoeuvre. Rescale $\tilde{g}$ to

$$\breve{g} := \left(\frac{1+L}{2}\right)^4 \tilde{g}. \tag{10.123}$$

[583] See Beig (1980), Kennefick & Ó Murchadha (1995), or Schoen (2009), Lecture 11. One needs two identities, cf. Wiki's *List of formulas in Riemannian geometry*. First, $\hat{R}_{ij} = \tilde{R}_{ij} - L^{-1}\tilde{\nabla}_i\tilde{\nabla}_j L + 2L^{-2}\tilde{\nabla}_i L\tilde{\nabla}_j L - L^{-1}\tilde{\Delta} L \cdot \tilde{g}_{ij}$, valid in $d = 3$. Using (8.102) and $L = \exp(U)$, this gives $\hat{R}_{ij} = 2\partial_i U \partial_j U$. Second, $\tilde{\Delta} f = e^{2U}(\hat{\Delta} - \hat{g}^{ij}\partial_i U \partial_i) f$ for any function f, also in $d = 3$, so taking $f = L = \exp(U)$ and using (8.102) gives $\hat{\Delta} U = 0$.

[584] This was noted at least as early as Synge (1960), §VIII.4. We learnt it from Cederbaum (2019), Lecture 1.

The Riemannian manifold $(\Sigma, \breve{g})$ has vanishing Ricci scalar and asymptotic mass, i.e.

$$\breve{R} = 0; \qquad \Pi^0(\breve{g}) = 0, \tag{10.124}$$

where Π^0 is defined by (8.103). The first property follows from a simple computation.[585] The second follows from the asymptotics (10.118), noting that Π^0 is determined by the $1/r$ term (cf. the computation of the Schwarzschild case in footnote 364). Since

$$\left(\frac{1+L}{2}\right)^4 = 1 - \frac{2m}{r} + O\left(\frac{1}{r^2}\right); \qquad \tilde{g}_{ij} = \left(1 + \frac{2m}{r}\right)\delta_{ij} + O\left(\frac{1}{r^2}\right), \tag{10.125}$$

the m/r terms cancel in the product $\breve{g}$. We would now like to invoke the second part of the positive mass theorem 8.5 in order to infer that $(\Sigma, \breve{g})$ is isometric to Euclidean space $(\mathbb{R}^3, \delta)$, so that $(\Sigma, \tilde{g})$ is conformally flat. But this does not work since Σ is not a manifold but a manifold with boundary, on top of which (and for that reason) it is not complete.

To remedy this, we perform a trick.[586] First as a manifold, we form the "double"

$$\Sigma_d = \Sigma \cup_{\partial\Sigma} \Sigma, \tag{10.126}$$

with metric $\tilde{g}_d$ given as the original one $\tilde{g}$ on both copies of Σ including their common boundary. The function L, though, is extended to a function L_d on Σ_d defined as L on one copy of Σ and as $-L$ on the other; this can be done continuously (though not smoothly) since $L = 0$ on $\partial\Sigma$ (and also the metric $\tilde{g}_d$ is no longer smooth on the boundary).

Now rescale $\tilde{g}_d$ through (10.123), leading to a Riemannian manifold $(\Sigma_d, \breve{g}_d)$ without boundary. The end where $L \to 1$ of course remains asymptotically flat, but because of the conformal transformation (10.123) the end where $L \to -1$ can be compactified by adding a single point (which for the metric $\tilde{g}$ would have been a two-sphere at infinity).[587] The ensuing Riemannian manifold $(\dot{\Sigma}_d, \breve{g}_d)$ is complete (basically since one end is asymptotically flat and the other end has been compactified). The computations that imply (10.124) also work for $(\dot{\Sigma}_d, \breve{g}_d)$, which, then, satisfies the hypotheses of the second part of Theorem 8.5 and hence is isometric to $(\mathbb{R}^3, \delta)$. Consequently, $(\dot{\Sigma}_d, \tilde{g}_d)$ is conformally flat, but since this is a local property we conclude that $(\Sigma, \tilde{g})$ is conformally flat.

3. *The spatial metric $\tilde{g}_S$ is spherically symmetric.* This is clear for $\tilde{g}_S$. For our general metric $\tilde{g}$ (re)constructed so far, spherical symmetry follows from conformal flatness in the situation where the Ricci tensor is given by (8.100). The proof uses the fact that $dL \neq 0$ and that the level sets $L =$ constant are (topologically) two-spheres, which in turn follows from from (10.118).[588] Up to the boundary where $L = 0$, the space Σ of Definition 8.4.1

[585] For a conformal transformation $\breve{g} = \Omega^2\tilde{g}$ we have $\breve{R} = \Omega^{-2}\tilde{R} - 4\Omega^{-3}\tilde{\Delta}\Omega + 2\Omega^{-4}\tilde{g}^{ij}\partial_i\Omega\partial_j\Omega$. Taking $\Omega = (1+L)^2/4$ and using (8.102) gives $\breve{R} = 0$.

[586] For any manifold M with boundary ∂M, the manifold $M \cup_{\partial M} M$ is topologically defined as $(M \sqcup M)/\sim$, where $x \sim y$ iff $x, y \in \partial M$ and $x = y$, with a unique smooth structure making this space a manifold (without boundary). More generally, given two manifolds with boundary M_1 and M_2 with a diffeomorphism $f : \partial M_1 \to \partial M_2$, one may define $M_1 \cup_f M_2$ as $(M_1 \sqcup M_2)/\sim$, where $x \sim y$ iff $x \in \partial M_1$ and $y = f(x) \in \partial M_2$. The existence of a smooth structure on this quotient derives from the ***collar neighbourhood theorem***, which states that for any manifold M with boundary ∂M there is a neighbourhood U of ∂M in M and an associated diffeomorphism $\psi : U \stackrel{\cong}{\to} \partial M \times [0,1)$.

[587] This requires much more detailed arguments, see Lemma 3 in Bunting & Masood-ul-Alam (1987).

[588] This requires some elliptic PDE theory and Morse theory. See Theorem 1 in Künzle (1971), which goes back to Lichnerowicz (1955), §78: an asymptotically flat space is either flat (which corresponds to the case $m = 0$), or, if $m \neq 0$, has $dL \neq 0$ throughout, with level sets $\cong S^2$. The maximum principle for elliptic PDEs (or the last claim of Theorem 8.5) gives the first claim, whereas the absence of points where $dL = 0$ makes all level sets homeomorphic to those near $r \to \infty$. From Definition 8.4 and the first entry in (10.118), these level sets are two-spheres.

is then foliated by the level sets of L, and one may set up a calculus à la (6.10) - (6.15), but in one dimension lower. Indicating this by the use of spatial indices, and writing

$$W := \tilde{\nabla}^i L \tilde{\nabla}_i L, \tag{10.127}$$

a computation shows that the Cotton tensor squared is given by

$$C_{ijk}C^{ijk} = L^{-4}W^2(8\sigma_{ij}\sigma^{ij} + W^{-2}h^{ij}\partial^i W \partial_j W), \tag{10.128}$$

where $h_{ij} = \tilde{g}_{ij} - n_i n_j$ is the projection onto the level sets, in terms of their normal $\vec{n}$. Thus conformal invariance, i.e. $C_{ijk} = 0$, enforces $\sigma_{ij} = 0$ and $h^{ij}\partial_j W = 0$. This makes the level sets, already known to be two-spheres topologically, also two-spheres metrically (i.e. with the usual $SO(3)$-invariant metric Ω), so that Σ is spherically symmetric.[589]

We have shown (at least in outline) that the Riemannian manifold with boundary $(\Sigma, \tilde{g})$ defined through the assumptions in Theorem 10.25 plus Definition 8.4 is spherically symmetric. Using Birkhoff's theorem in the simple case where staticity has already been assumed, the spatial Schwarzschild metric (10.117) and then the full one (9.15) via (8.96) then follow. The case $m \leq 0$ is excluded by the assumptions in Theorem 10.25, since L has no zeros in that case. □

Short of Birkhoff's, this is the simplest uniqueness theorem for black holes! Similar reasoning shows that the subcritical Reissner–Nordström metric (i.e. $0 < |e| < m$) is the unique (exterior) static "electrovac" black hole space-time with smooth *non-degenerate* event horizon and vanishing magnetic charge (where 'non-degenerate' means nonzero surface gravity).[590] However, the degenerate case ($|e| = m > 0$) is not unique! Consider the Reissner–Nordström metric (9.88) for this case, given by the usual static metric (8.96), now with spatial part

$$\tilde{g}_{RNd} = L(r)^{-2}dr^2 + r^2(d\theta^2 + \sin^2\theta d\varphi^2); \qquad L(r) = 1 - m/r. \tag{10.129}$$

Compared with the Schwarschild metric (10.117), we have $1 - m/r$ instead of $\sqrt{1-2m/r}$. The coordinate transformation $\rho = r - m$ and then $\rho \rightsquigarrow r$ turns the total space-time metric into

$$g_{MP} = -\frac{dt^2}{U^2} + U^2(dr^2 + r^2 d\Omega), \tag{10.130}$$

where $r > 0$ and

$$U := 1 + m/r. \tag{10.131}$$

Clearly, U solves the (flat space) Laplace equation

$$\Delta U = 0, \tag{10.132}$$

The point is that (10.130) is a static solution to the Einstein–Maxwell equations for *arbitrary* (static) solutions to (10.132), provided the electromagnetic field potential is taken to be

$$A = U^{-1}dt. \tag{10.133}$$

As such, (10.130) is called a ***Majumdar–Papapetrou metric***. For example, one may take

$$U = 1 + \sum_{i=1}^{N} \frac{e_i}{|\vec{x} - \vec{y}_i|}, \tag{10.134}$$

[589] See Theorem 2 in Künzle (1971) or Corollary 9.3 in Heusler (1996).

[590] See Israel, 1968, Masood-ul-Alam (1992), and Heusler (1996), §9.3, and further references in the latter.

where $(\vec{y}_1,\dots,\vec{y}_N) =: Y$ is any finite set of points in $\mathbb{R}^3$, so that g_{MP} is defined on

$$M_{MP} := \mathbb{R}\times\mathbb{R}^3\backslash(\vec{y}_1,\dots,\vec{y}_N). \tag{10.135}$$

It turns out that the event horizon H_E^+ equals Y, and that e_i is the charge at the puncture $\vec{y}_i$.[591]

This leads to a generalization of Theorem 10.25, which like most uniqueness theorems describes the domain of outer communication (DOC). Apart from its axiomatic definition (10.83) in the presence of $\mathscr{I}$, this domain can also be defined for a stationary asymptotically flat space-time (M,g), with asymptotically timelike Killing vector X $(= \partial_t)$. Just calling it D, we put

$$D := I^-(M^{\text{ext}})\cap I^+(M^{\text{ext}}); \qquad M^{\text{ext}} := \cup_{t\in\mathbb{R}}\varphi_t(\Sigma^{\text{ext}}), \tag{10.136}$$

where Σ^{ext} is one of the ends of M as in Definition 8.4.1, and φ_t is the flow of X (assumed complete). For example, in Kruskal space-time with Σ^{ext} "to the right", D corresponds to region I. Similarly, we may define black and white hole regions and their event horizons by

$$B^\pm := M\backslash I^\mp(M^{\text{ext}}); \tag{10.137}$$

$$H_E^\pm := \partial B^\pm = \partial I^\mp(M^{\text{ext}}). \tag{10.138}$$

Recent uniqueness theorems assume that D is globally hyperbolic,[592] which is a "Penrosian" version of weak cosmic censorship, see the end of §10.4. In the static case, we then have:[593]

Theorem 10.26 *Let (M,g) be a static asymptotically flat electrovac space-time (i.e. solving the source-free Einstein–Maxwell equations) containing a connected acausal spacelike hypersurface Σ (cf. Definition 8.4.3) whose closure $\overline{\Sigma}$ is a topological manifold with boundary consisting (as a disjoint union) of a compact set K and finitely many ends Σ_α^{ext} (cf. Definition 8.4.1). If the* DOC *(D,g) is globally hyperbolic and $\partial\overline{\Sigma}\subset M\backslash D$, then Σ can have only one asymptotic end, and:*

- *If the event horizon H_E^+ defined in (10.138) is connected, then the* DOC *of the unique end Σ^{ext} is isometric to the* DOC *of a Reissner–Nordström space-time with $0\leq|e|\leq m\neq 0$.*
- *If the event horizon H_E^+ is not connected, then the* DOC *of Σ^{ext} is isometric to the* DOC *of a Majumdar–Papapetrou space-time with $N\geq 2$.*

In particular, in the vacuum case one recovers the DOC *of Schwarzschild space-time with $m>0$.*

Here Reissner–Nordström includes $e=0$, i.e., Schwarzschild with $m>0$. The cases $|e|>m>0$ and $e=0>m$ are excluded because they lack event horizons (and have $D=M$, which is not globally hyperbolic, cf. §10.6). The Minkowski case $e=m=0$ has no event horizon either.

[591] See Chruściel (2020), §4.7. The charge is defined by $-(4\pi)^{-1}\int_{S_i^2}*F$, where S_i^2 is some two-sphere around $\vec{y}_i$.

[592] Global hyperbolicity of D is clearly a necessary condition for the theorems (as the space-times in their conclusions satisfy it). Galloway (1995) showed that global hyperbolicity of D plus the null energy condition (which is satisfied here) imply that D is is simply connected, which is needed for the proof of Theorem 10.26 (which we omit). Global hyperbolicity of D is also needed to underpin the heavy PDE analysis in all proofs (Chruściel & Lopes Costa, 2008). The original assumption used in this context by Hawking and others in the 1970s was *strong asymptotic (future) predictability*, see footnote 638.

[593] Theorem 10.26 is Theorem 3.1 in Chruściel, Lopes Costa, & Heusler (2012), which as the authors explain is the culmination of a long development starting with Israel's theorem 10.25 (and we would say: with Birkhoff's). The implication that Σ can then only have one asymptotically flat end is part of Theorem 3.3.1 in Chruściel (2020), originally due to Chruściel & Wald (1994a). This follows from topological censorship, see footnote 601 in §10.10.

10.10 Black hole uniqueness theorems: Stationary case

Passing from the static to the stationary case, a natural generalization of Theorem 10.26 would be that *the domain of outer communication* (D,g) *is isometric to the* DOC *of a Kerr–Newman space-time (with* $0 \leq a^2+e^2 \leq m^2$*),* at least if the event horizon is connected.[594] Unfortunately, this has only been proved under considerably stronger assumptions, namely:[595]

1. Those of Theorem 10.26, of course replacing *static* by *stationary*;[596]
2. Connectedness and non-degeneracy of the horizon H_E^+ (i.e. surface gravity $\kappa \neq 0$).[597]
3. $\partial\overline{\Sigma} \subset \partial D \cap I^+(D)$, such that $\partial\overline{\Sigma}$ intersects each null generator of $\partial D \cap I^+(D)$ once.[598]
4. Analyticity of the space-time metric g.[599]

Theorem 10.27 *Let* (M,g) *be a stationary asymptotically flat electrovac space-time satisfying 1–4. Then* (D,g) *is isometric to the* DOC *of a Kerr–Newman space-time with* $0 \leq a^2+e^2 < m^2$.

Thus (M,g) is characterized by just three numbers (m,a,e) and hence this is the ultimate "no-hair theorem". An important stepping stone from the static to the stationary case is ***Hawking's rigidity theorem***, which is very interesting by itself and explains the coincidence of event horizons and Killing horizons in the Schwarzschild, Reissner–Nordström, and Kerr metrics.[600]

Theorem 10.28 *Under the assumptions of Theorem 10.27, either the asymptotically timelike Killing vector field* X *defining stationarity is tangent to the event horizon* H_E^+*, or the isometry group of* (M,g) *contains* $\mathbb{R}\times U(1)$*, where* $U(1)$ *acts via spatial rotations, and there is another vector field* Y *that is a linear combination of* X *and the generator* ∂_φ *of the* $U(1)$ *isometries, for which* H_E^+ *is a Killing horizon. Either way, the event horizon* H_E^+ *is also a Killing horizon.*

Although the first option describes the situation for the Schwarzschild metric, see Theorem 9.1, and the second one for the Kerr metric, see Theorem 9.3 and eq. (9.143), in general it seems quite mysterious where the axial symmetry should come from. This much we will explain. A key lemma for Hawking's rigidity theorem and a very important result in its own right is:[601]

[594] There are various candidates for space-times with multiple rotating black holes generalizing the Majumdar–Papapetrou metrics to the stationary case, none of which are well understood. See e.g. Weinstein (1996) as well as numerous physics papers, partly reviewed in Chruściel, Lopes Costa, & Heusler (2012), §3.2.2.

[595] Theorem 10.27 is due to Chruściel & Lopes Costa (2008); see also Lopes Costa (2010). There are many other stationary axisymmetric solutions (Stephani *et al.*, 2003, chapters 19–21) but these are either not asymptotically flat or have non-globally hyperbolic DOC and hence lack an event horizon and fail weak cosmic censorship.

[596] Completeness of the Killing field X in charge of stationarity is included as part of the definition of the latter.

[597] One may drop non-degeneracy at the expense of *assuming* $U(1)$-invariance (Chruściel & Nguyen, 2010).

[598] By (10.136) and Proposition 10.16.1, $\partial D \cap I^+(D)$ is ruled by lightlike geodesics. This assumption is technical and is explained in detail in Lopes Costa (2010). Roughly speaking, the specific cross-section of the horizon given by intersection with $\partial\overline{\Sigma}$ must hit its null generators once. This is clearly the case for Kerr–Newman.

[599] This is the most undesirable hypothesis, required for Theorem 10.28. See also footnote 604.

[600] The original version is in Hawking (1972) and Hawking & Ellis (1973), Proposition 9.3.6. Theorem 10.28 is like Theorem 5.1 in Chruściel (1996), based on Chruściel (1997). See also Friedrich, Rácz, & Wald (1999).

[601] This proposition goes back to Hawking (1972) and Hawking & Ellis (1973), §9.3, with dubious proof. The approach via topological censorship, introduced by Friedman, Schleich, & Witt (1993), is due to Chruściel & Wald (1994a), Galloway (1995), Browdy & Galloway (1995), and Jacobson & Venkataramani (1995). For the topological singularity theorem of Gannon (1995) and Lee (1976) see footnote 311. Overall, see §3.3 in Chruściel

Proposition 10.29 *Let (M,g) be a stationary asymptotically flat space-time satisfying the null energy condition, as well as weak cosmic censorship in the sense that its* DOC *D is globally hyperbolic. Accordingly, let Σ be a Cauchy surface in D, with $\Sigma^{ext} \cong (0,\infty)\times S^2$. Then:*

1. *The domain of outer communication D is simply connected.*
2. *If the closure $\overline{\Sigma}$ of Σ in M intersects the future event horizon H_E^+ in a compact set K, and H_E^+ is connected, then $K \cong S^2$ and hence $H_E^+ \cong \mathbb{R}\times S^2$ (both meant topologically).*

Both results are very deep and as usual by now, we can only sketch the arguments. The inference from global hyperbolicity to simple connectedness is essentially the principle of ***topological censorship***, which is proved (by contradiction) by combining ideas from Corollary 10.18 and the topological singularity theorem, which all go back to Penrose's singularity theorem. The second claim follows from the first, combined with a result from differential topology:[602]

Lemma 10.30 *If N is a compact simply connected 3-manifold with non-empty boundary ∂N, then all connected components of ∂N must be diffeomorphic to two-spheres S^2.*

The simplest example is the three-ball B^3 with $\partial B^3 \cong S^2$. In the case at hand, Theorem 5.33 gives $D \cong \mathbb{R}\times\Sigma$ and hence Σ is simply connected by part 1 of Proposition 10.29. Since (M,g) is asymptotically flat we can cut off Σ at some large radius r, giving the N of the lemma. The component of ∂N at the asymptotically flat end is $\cong S^2$ (this much is clear even without the lemma, which of course confirms it), and the other is $\Sigma\cap H_E^+ \cong S^2$ (from Lemma 10.30). □

Towards Theorem 10.28, by stationarity of the metric, the event horizon (which is defined by the causal structure and hence by the metric) is invariant under the flow of X (which by definition of a Killing vector field consists of isometries), and hence X is tangent to H_E^+. Thus $X = L - Z$ on H_E^+, where L is tangent to the null generators of H_E^+, and Z is tangent to the spacelike two-spheres S^2 of Proposition 10.29.2. Let $\check{g} = \sum_{i,j=1}^2 \check{g}_{ij}dx^i dx^j$ be the Riemannian metric on S^2 (so far meant topologically, rather than metrically), and write $L = d/ds$ and $Z = \sum_{i=1}^2 Z^i\partial_i$. Using $\check{g}_{is} = \check{g}_{sj} = 0$ (expressing orthogonality of L to the null hypersurface H_E^+, to which L is simultaneously tangent!), the Killing equation $\mathscr{L}_X\check{g}_{ij} = 0$, given by (2.94), comes down to

$$\mathscr{L}_Z\check{g} + \partial_s\check{g}_{ij}dx^i dx^j = 0; \qquad \partial_s Z^i = 0. \tag{10.139}$$

Now the conceptual key to Hawking's argument is that, (possibly) apart from the isometries generated by the "stationarity" Killing vector field X, at least the intrinsic geometry of the horizon of a stationary black hole, as determined by $\check{g}$, is also invariant under the flow of L, i.e. along its null generators. This will follow from the arguments leading to (10.170) below, which give $k_{\mu\nu} = 0$ *on the horizon*. Eq. (8.23) then gives $\partial_s\check{g}_{ij} = 0$, so that (10.139) becomes

$$\mathscr{L}_Z\check{g} = 0; \qquad \partial_s Z^i = 0. \tag{10.140}$$

(2020). As noted at the end of §7.3, the null energy condition required in Proposition 10.29 is satisfied by electrovac space-times, so we need not assume it separately in Theorem 10.27. Finally, Proposition 10.29 speaks of "the" DOC, since by footnote 593 there can only be one asymptotically flat end in M. See also footnote 593.

[602] This is Lemma 4.9 in Hempel (1976). A simpler proof, kindly provided by my colleague Ioan Mărcuț, uses a long exact sequence in de Rham cohomology, viz. $0 \to H^0(\partial N) \to H^2(N) \to H^1(N) \to H^1(\partial N) \to H^1(N) \to \cdots$, valid in $d = 3$. By assumption, $H^1(N) = 0$, which gives $0 \to \mathbb{R} \to H^0(\partial N) \to H^2(N) \to 0$ as well as $H^1(\partial N) = 0$. The former gives $\dim(H^2(N)) + 1$ components of ∂N and the latter makes each of these diffeomorphic to S^2.

Hence the vector field Z, so far defined only on H_E^+, is independent of s and generates (the same) isometries on each spacelike two-sphere within H_E^+ that is orthogonal to L, i.e. to the null generators of the horizon. Now there are clearly two mutually exclusive possibilities:

- Either $Z = 0$, in which case $X = L$ is tangent to the horizon, which thereby becomes a Killing horizon with respect to X. By highly nontrivial further arguments going under the name ***staticity theorem***,[603] the stationary case is eventually reduced to the static case.
- Or $Z \neq 0$, in which case $X + Z = L$ is tangent to the horizon, making it a Killing horizon for a new Killing vector field. By Lemma 10.31 below, Z has periodic orbits on the horizon, and by (10.140) the vector fields Z and X commute. This eventually leads to the factorization $\mathbb{R} \times U(1)$ of (asymptotic) time-evolution and rotation. As Hawking suggested, the extension of Z and hence of the $U(1)$ symmetry it generates off the horizon to all of M can be done via analyticity of the metric, which is why this was assumed.[604]

Lemma 10.31 *If a (Riemannian) metric $\check{g}$ of a two-sphere S^2 (seen as a manifold only) admits a nonzero Killing vector field Z, then the orbits of the flow of Z are closed (i.e. periodic).*[605]

Proof.[606] First, like any vector field on a compact manifold, Z is complete and hence has a globally defined flow $\psi : \mathbb{R} \times S^2 \to S^2$; as usual we write $\psi_t(x) = \psi(t,x)$, with $\psi_t : S^2 \to S^2$. By the "hairy ball" theorem,[607] Z has a zero on S^2, say at z (for $Z = \partial_\psi$ on S^2 with the usual metric, z would be the north pole or the south pole). The tangent map $T_z\psi_t$ ($= (\psi_t)_*$ at z) then maps T_zS^2 to itself (for each $t \in \mathbb{R}$), and since each ψ_t is an isometry, $T_z\psi_t$ is an isometry of $\check{g}_z$. Identifying $T_zS^2 \cong \mathbb{R}^2$ through the choice of an orthonormal basis, we have $T_z\psi_t \in SO(2)$, and more precisely, $T_z\psi_t = \exp(tA)$ for some A in the Lie algebra of $SO(2)$. Consequently, $T_z\psi_T = \mathrm{id}$ for some (smallest) $0 < T < \infty$; we may normalize Z such that $T = 2\pi$. Furthermore,

$$\psi_t(\exp_z(V)) = \exp_z(T_z\psi_t(V)). \tag{10.141}$$

By Hopf–Rinow, the map $\exp_z : T_zS^2 \to S^2$ is surjective, and hence $\psi_T(x) = x$ for any $x \in S^2$. □

[603] Such a theorem shows that under the assumptions of Theorem 10.27, where in addition (M,g) is not rotating, X is hypersurface orthogonal (Sudarsky & Wald, 2002, 2003; Chruściel & Wald, 1994b; Heusler, 1996, §8.2).

[604] See Hawking & Ellis (1973), Proposition 9.3.6, Chruściel (1996), Lemma 5.2 and Heusler (1996), §8.1. Stationary vacuum solutions can only be *proved* to be analytic (i.e. $g_{\mu\nu}$ has a convergent power series expansion) where X is timelike (Müller zum Hagen, 1970ab), i.e. in the DOC. Attempts to remove analyticity of the metric from the assumptions of Theorems 10.27 and 10.28 have led to a program (still in progress) called ***Kerr rigidity***. This aims at a different version of the black hole uniqueness theorems, where in compensation for weakening the assumptions along the above lines, one also has to strengthen them, in that (so far in the vacuum setting) one tries to show that at least stationary solutions to Einstein's equations that are close to Kerr, in the DOC actually coincide with Kerr. See Alexakis, Ionescu, & Klainerman (2014) and Ionescu & Klainerman (2015). Kerr *rigidity* is to be distinguished from ***Kerr stability***, which is the conjecture that generic perturbations of the initial data for the Kerr metric lead to a MGHD that is close to the original one (at least in the DOC). This would generalize the paradigmatic theorem on the stability of Minkowski space-time (Christodoulou & Klainerman, 1993; Lindblad & Rodnianski, 2010) to black holes. There is numerical evidence for this (Zilhão et al., 2014), and recent mathematical results prove it for $a = 0$, i.e. Schwarzschild (Dafermos, Holzegel, Rodnianski, & Taylor, 2021), and for small a, both for cosmological constant $\lambda = 0$ (Klainerman & Szeftel, 2021) and $\lambda > 0$ (Hintz & Vasy, 2018).

[605] There is a smallest period T_0 of which all other periods are integral multiples. First, by the *period bounding lemma* (stating that the non-zero periods of a vector field on a compact manifold are bounded from below) there are only finitely many periods. Second, by the proof of Lemma 10.30, each periods equals T/n_i, for some $n_i \in \mathbb{N}$. Taking $n_0 = \mathrm{LCM}(n_1,\ldots,n_k)$, where T/n_i are the periods, it follows that each period is a multiple of T/n_0.

[606] This proof was again kindly provided by my colleague Ioan Mărcuț. The lemma fails if S^2 is replaced by for example the 2-torus (Kronecker foliation) or the three-sphere (Reeb foliation), cf. Moerdijk & Mrcun (2003), §1.1.

[607] This theorem is often attributed to Brouwer (who generalized it to S^{2n}), but it goes back to Poincaré (1885).

At this stage we know that a space-time satisfying the assumptions of Theorem 10.27 is both *stationary* and *axisymmetric*, in that its isometry group contains $\mathbb{R}\times U(1)$, where $\mathbb{R}$ at least in the DOC gives timelike transformations, whereas $U(1)$ gives spatial rotations around a symmetry axis (which consists of all points where the Killing vector field Z generating these rotations vanishes).[608] The next step is the ***circularity theorem*** to the effect that the distribution orthogonal to X and Z is integrable,[609] so that, roughly speaking, in suitable coordinates the 2-surfaces generated by the $\mathbb{R}\times U(1)$ action (which in the DOC have constant r and θ) are orthogonal to 2-surfaces having constant t and φ. The assumptions in this theorem are automatically satisfied when the Ricci tensor vanishes, so for simplicity we restrict ourselves to the vacuum case.[610]

The circularity theorem brings the metric into the so-called ***Papapetrou form***,

$$g=-\rho^2e^{2\lambda}dt^2+e^{-2\lambda}(d\varphi+Adt)^2+e^{2\lambda}e^{2h}(d\rho^2+dz^2), \tag{10.142}$$

in coordinates (t,ρ,z,φ) resembling the usual cylindrical coordinates $(x=\rho\cos\varphi, y=\rho\sin\varphi, z)$, where the functions λ, A, and h only depend on (ρ,z). In terms of the $2d$ Riemannian manifold $(\overline{\Sigma},\overline{g})$ defined by $\overline{\Sigma}=\mathbb{R}\times(0,\infty)$, coordinatized by $\rho>0$ and $z\in\mathbb{R}$, and $\overline{g}=e^{2h}(d\rho^2+dz^2)$, solving the vacuum Einstein equations $R_{\mu\nu}=0$ then comes down to solving the elliptic PDE

$$X\overline{\nabla}_i(\rho\overline{\nabla}^iE)+\rho\overline{\nabla}_iE\overline{\nabla}^iE=0, \tag{10.143}$$

called the (vacuum) ***Ernst equation***, for the complex ***Ernst potential*** $E=-X+iY$, with $X>0$. Here $i=1,2$, and $\overline{\nabla}$ is the covariant derivative defined by the $2d$ metric $\overline{g}_{ij}$. Namely, if we know E we find λ from $X=\exp(-2\lambda)$, whereas A and h come from solving the first-order PDEs

$$\partial_\rho A=\frac{\rho}{X^2}\partial_zY; \qquad \partial_\rho h=\frac{\rho}{X^2}(\partial_\rho E\partial_\rho\overline{E}-\partial_zE\partial_z\overline{E}); \tag{10.144}$$

$$\partial_z A=-\frac{\rho}{X^2}\partial_\rho Y; \qquad \partial_z h=\tfrac{1}{4}\frac{\rho}{X^2}(\partial_\rho E\partial_z\overline{E}+\partial_zE\partial_\rho\overline{E}). \tag{10.145}$$

Eq. (10.143) is subject to boundary conditions dictated by the assumptions in Theorem 10.27,[611] and the last difficult part of the proof of Theorem 10.27 is to show that these conditions precisely allow the Kerr metric (or, in the electrovac case, where (10.143) has extra terms, the Kerr–Newman metric) and no other solutions. This has been done in at least four different ways, none of which is easily explained.[612] The general point, though, is that through stationarity and its consequence axisymmetry, the vacuum Einstein equations have been reduced to a $2d$ elliptic boundary value problem, which can be completely controlled and gives the desired uniqueness.

These results are very impressive and should suffice for stationary (i.e. long-term) astrophysical situations. However, from a theoretical point of view it should be stressed that couplings to other forms of matter than electromagentism typically do give "hair" to black holes.[613]

[608] This set is non-empty by Lemma 10.30 and Poincaré's hairy ball theorem used in the proof of Lemma 10.31.

[609] Compare with Lemma 10.24.2, which in the spherically symmetric case gives 2-surfaces with constant r and t that are orthogonal to 2-surfaces with constant θ and φ, viz. the leaves of the S^2- foliation of Lemma 10.24.1.

[610] For all that follows, see Carter (1979, 1986), Heusler (1996), and Chruściel, Lopes Costa, & Heusler (2012).

[611] In order to resolve the ambiguity that $\rho=0$ both at the horizon and at the zeros of the rotation generator Z, these boundary conditions are usually stated in terms coordinates (x,y) instead of (ρ,z), constrained by so as to lie in the semi-strip $x>C:=m-2\Omega J$ and $-1<y<1$, and defined by $\rho=\sqrt{(x^2-C^2)(1-y^2)}$ and $z=xy$. In terms of these, the $2d$ metric becomes $d\rho^2+dz^2=dx^2\cdot(x^2-y^2)/(x^2-C^2)+dy^2/(1-y^2))$, and $x\to C$ is the horizon, whereas $y\to\pm1$ is the symmetry axis. See Carter (1986), p. 106, or Heusler (1996), p. 55. For example, asymptotic flatness gives boundary conditions as $x\to\infty$, namely $x^{-2}X=(1-y^2)(1+O(1/x))$ and $Y=2Jx(3-y^2)+O(1/x)$, where J is a constant. As $y\to\pm1$, we have X, ∂_xY, and ∂_yY all $O(1-y^2)$, and $\partial_yX=c+O(y^2-1)$ for some $c>0$.

[612] These are reviewed, with references to the original literature, in Chruściel, Lopes Costa, & Heusler (2012).

[613] See e.g. Volkov, 2018, and references therein, as well as, again, Chruściel, Lopes Costa, & Heusler (2012).

10.11 The Penrose inequality

The final state conjecture mentioned in §10.9 has an interesting and testable consequence known as the ***Penrose inequality***.[614] Because of its paramount role in the final state conjecture, the Penrose inequality is often seen as a test of weak cosmic censorship. The logic seems to be:

final state conjecture $\Rightarrow$ *weak cosmic censorship* $\Rightarrow$ *Penrose inequality,*

where at least Penrose himself seemed clearly interested in the contrapositive implication

violation of Penrose inequality, $\Rightarrow$ *violation of weak cosmic censorship.*

To motivate the inequality, let us first define the ***area*** A_K of a Kerr black hole in the regime $0 < |a| \leq m$, where weak cosmic censorship holds (see Theorem 9.3 in §9.7 and §10.6), as the area of its (future) event horizon H_E^+ at some fixed value v_0 of the lightlike coordinate $v \equiv v_+$ defined above (9.136). Since $r = r_+$ at this horizon, one may also say that t is fixed, and hence that A_K is the area of the intersection of H_E^+ with some spacelike "wannabe" (i.e. partial) Cauchy surface Σ. Since the metric is stationary, this area is in fact independent of v of t or Σ. We obtain

$$\begin{aligned} A_K &:= \int_0^\pi d\theta \int_0^{2\pi} d\varphi \sqrt{\det h(v_0, r_+, \theta, \varphi)} = \int_0^\pi d\theta \int_0^{2\pi} d\varphi \sqrt{\Sigma(r_+)} \sin\theta \\ &= \int_0^\pi d\theta \int_0^{2\pi} d\varphi \, (r_+^2 + a^2) \sin\theta = 4\pi(r_+^2 + a^2) = 8\pi(m^2 + m\sqrt{m^2 - a^2}), \end{aligned} \tag{10.146}$$

where h is the metric on the set $H_E^+ \cap \{v = v_0\}$ induced by the Kerr metric (9.137). Here we used (9.113), of which only the first term remains, since $\Delta = 0$ at $r = r_+$, cf. (9.117). For the Schwarzschild black hole, in which $a = 0$, with $r_S = 2m$ this simply gives

$$A_S = 4\pi r_S^2 = 16\pi m^2. \tag{10.147}$$

For the Kerr metric, still assuming $0 < |a| \leq m$ and hence weak cosmic censorship, we have

$$A_K \leq 16\pi m^2. \tag{10.148}$$

This fact about Kerr space-time is the key to the Penrose inequality. It gives a positive lower bound on the (asymptotic) mass of a black hole in terms of the area of some spatial cross-section of its event horizon, and the inequality is saturated by the Schwarzschild metric.

Suppose the Kerr black hole is the final state of a gravitational collapse process. At some earlier time t, captured by a spacelike hypersurface $\Sigma_t \equiv \Sigma$, assuming weak cosmic censorship, an event horizon has formed with spatial or cross-sectional area $A_t \equiv A_\Sigma$ at time t, that is,

$$A_\Sigma := \text{Area}(H_E^+ \cap \Sigma). \tag{10.149}$$

Here H_E^+ is the event horizon of the space-time of the collapsing matter as defined in (10.79). By Hawking's area law (10.160), to be discussed in the next section, the area can only *increase* during the collapse process, so that $A_\Sigma \leq A_K$, since $A_K \equiv A_\infty$ is now seen as the horizon area at $t = \infty$, where the collapse has been completed and a stationary Kerr black hole has been formed.

[614] The original source is Penrose (1973), to be recommended also for its figures. Penrose describes weak cosmic censorship and the final state conjecture, both of which rank among his most visionary contributions to GR, as 'the *establishment* viewpoint', and sees his inequality as 'an attempt to derive a contradiction with this viewpoint.' The subject can be traced through the reviews by Bray (2002), Bray & Chruściel (2004), Mars (2009), and Lee (2019).

On the other hand, the mass $m_t \equiv m_\Sigma$ of the black hole at time t (the "now" of Σ) can only *decrease* through gravitational radiation, i.e. $m_t \geq m \equiv m_\infty$. Thus (10.148) gives $A_t \leq 16\pi m_t^2$, or

$$A_\Sigma \leq 16\pi m_\Sigma^2, \tag{10.150}$$

where, in the presence of possible asymptotic momentum, the asymptotic mass is defined by

$$m_\Sigma := \sqrt{(\Pi^0)^2 - \|\vec{\Pi}\|^2}. \tag{10.151}$$

Here Π^0 and $\vec{\Pi}$ are defined by (8.103) and (8.109), respectively, and the spacelike hypersurface Σ is supposed to carry initial data $(\tilde{g}, \tilde{k})$ satisfying the inequality (8.110). Since this implies

$$\Pi^0 \geq \|\vec{\Pi}\| \tag{10.152}$$

by the (generalized) positive mass theorem, the mass m_Σ in (10.151) is well defined, and (10.151) simply reflects the basic formula $p^0 = \sqrt{|\vec{p}|^2 + m^2}$ from relativistic mechanics. Thus the assumption (8.110) on the initial data will be made throughout this section.[615] This is not as threatening as it sounds, since in the main case of interest, i.e. the static case, one has $\tilde{k} = 0$ and hence (8.110) just comes down to non-negative scalar curvature, i.e. $\tilde{R} \geq 0$. See below.

Eq. (10.150), then, is a first version of the Penrose inequality. However, since the event horizon H_E^+ has the disadvantage explained at the end of §10.3, which makes A_Σ effectively uncomputable from the initial data on Σ, the meaning of the inequality must be modified. The idea is to replace A_Σ by some computable number $\tilde{A}_\Sigma$ resembling the area of a spatial cross-section of the event horizon, in such a way that $\tilde{A}_\Sigma \leq A_\Sigma$. The redefined Penrose inequality would then become $\tilde{A}_\Sigma \leq 16\pi m_\Sigma^2$, and although this is weaker than (10.150) and hence its *proof* would give less information than a proof of (10.150), a *violation* of the weaker version, i.e. $\tilde{A}_\Sigma > 16\pi m^2$, would still falsify cosmic censorship, as Penrose intended in Popperian spirit.

A natural candidate to replace the "absolute" event horizon (10.79) is the ***apparent horizon***. Its definition relies on the notion of an outer trapped surface, which is predicated on the possibility of defining an *outer* direction among the pair of lightlike vectors $(L, \underline{L})$ emanating from a closed spacelike surface S, as defined in §6.3. This can be done, for example, if the given space-time (M, g) has a non-compact spacelike (full or wannabe) Cauchy surface Σ and S is such that $S \subset \Sigma$ and $\Sigma \backslash S = U \sqcup V$ with $\overline{U}$ compact and $\overline{V}$ non-compact, so that S separates Σ into an inside part U and an outside part V. In that case, the outer lightlike vector field L is selected by $g(L, n) > 0$, where n is the outward normal to S within Σ (i.e. n points towards V). This applies if (M, g) is asymptotically flat (cf. Definition 8.4) and S is the boundary of a region that does not extend to the asymptotic end Σ^{ext} of Σ (we assume there is only one such end). *We only consider closed spacelike surfaces S with this property, which are simply called* **surfaces** *in what follows.*

In the presence of a preferred outer direction we write (L^+, L^-) for $(L, \underline{L})$, normalized to $g(L^+, L^-) = -2$ as in (6.58), and similarly write (θ^+, θ^-) for $(\theta, \underline{\theta})$. If $S \subset \Sigma \subset M$, with Σ a spacelike hypersurface, as above, as usual we write N for the fd normal to the embedding $\Sigma \hookrightarrow M$, normalized by $g(N, N) = -1$, and denote the corresponding extrinsic curvature by $\tilde{k}$. Furthermore, let n be the outward directed normal to the purely Riemannian embedding $S \hookrightarrow \Sigma$, with extrinsic curvature $\check{k}$. Generalizing (6.59) - (6.60), one has

$$L^\pm = N \pm n; \qquad \theta^\pm = \text{Tr}_S(\tilde{k}) \pm \text{Tr}_S(\check{k}), \tag{10.153}$$

[615] Eq. (8.110) follows if the constraints (8.65) - (8.67) as well as the dominant energy condition $E \geq \|\vec{P}\|$ hold. But since the matter content is not specified, the constraints are usually not imposed in the Penrose inequality.

where $= \mathrm{Tr}_S(\tilde{k})$ is the trace of the pull-back of $\tilde{k}$ to S under the embedding $S \hookrightarrow \Sigma$, and $\mathrm{Tr}_S(\check{k})$ might as well have been written $\mathrm{Tr}\,(\check{k})$ since $\check{k}$ was already defined on S in the first place.[616]

The following definition may look unnecessarily complicated, but that's the way it is:[617]

Definition 10.32 *In the above circumstances, a surface $S \subset \Sigma \subset M$ is:*

- **future outer trapped** *if $\theta^+ < 0$, cf. (6.74) and (6.96);*
- **weakly outer trapped** *if $\theta^+ \leq 0$,*
- **marginally outer trapped** *if $\theta^+ = 0$, in which case we call S an* MOTS.

1. *The* **outer trapped region outer trapped region** *$T^+_\Sigma \subset \Sigma$ is is the union of the interiors of all weakly outer trapped surfaces in Σ.*
2. *The* **apparent horizon** *of M within Σ is $A^+_\Sigma := \partial T^+_\Sigma$.*

In the asymptotically flat case, it can be shown that the apparent horizon is smooth and is an MOTS,[618] which by definition encloses all weakly outer trapped surfaces in Σ. For stationary black holes the apparent horizon A^+_Σ coincides with $H^+_E \cap \Sigma$, which therefore is an MOTS.

For example, for the Schwarzschild metric the property $\theta^+ = 0$ easily follows from eqs. (6.96), and (9.49), and (9.44). In fact, we also find $\theta^- = 0$, either by computation, or because

$$\theta^- = -\theta^+ \tag{10.154}$$

in static space-times. This follows from (10.153), since now $\tilde{k} = 0$ and hence $\theta^\pm = \pm \mathrm{Tr}_S(\check{k})$.

Proposition 10.33 *An* MOTS *$S \subset \Sigma \subset M$ in a static space-time (M, g) has lightlike expansions*

$$\theta^+ = \theta^- = 0. \tag{10.155}$$

In particular,[619] S is a minimal surface in the 3d Riemannian manifold Σ.

Proof. This follows immediately from the definitions and from eqs. (10.154) and (6.85). □

This proposition suggests that in general space-times MOTSs are Lorentzian analogues of minimal (hyper)surfaces in Riemannian geometry, which partly explains their enormous interest. More importantly, Proposition 10.33 is the key to the Riemannian Penrose inequality below.

We return to the general (i.e. not necessarily static) case. A slight variation of Corollary 10.18 shows that the apparent horizon lies within the event horizon.[620] But this does not mean that $\mathrm{Area}(A^+_\Sigma) \leq \mathrm{Area}(H^+_E \cap \Sigma)$, and hence, taking the left-hand side to be $\tilde{A}_\Sigma$ and the right-hand side as A_Σ, the desired inequality $\tilde{A}_\Sigma \leq A_\Sigma$ fails. This may be remedied as follows.[621]

---

[616] See e.g. Minguzzi (2019), §6.4, for a derivation of (10.153) and similar results.

[617] Hawking & Ellis (1973), §9.2, pp. 319–320, define the apparent horizon as the boundary of the outer trapped region, which they define as the set of all points $x \in \Sigma$ that lie on some outer trapped surface. However, this faces problems with the smoothness of the boundaries involved. See Andersson & Metzger (2009) and Chruściel (2020), §8.4. Andersson, Mars, & Simon (2008) and Galloway, Miao, & Schoen (2015) are references on MOTSs.

[618] See Andersson & Metzger (2009), Theorem 7.3.

[619] A ***minimal surface*** in a Riemannian manifold (locally) minimizes the volume functional, which is the case iff its mean extrinsic curvature vanishes. See e.g. Jost (2002), §3.6. Euclidean space $\mathbb{R}^3$ only has non-compact minimal surfaces; beside the (affine) planes, one has interesting examples like the *catenoid* and the *helicoid* (see Wiki).

[620] See Proposition 9.2.8 in Hawking & Ellis (1973) and Theorem 3.3.18 in Chruściel (2020).

[621] One could sharpen this definition to make $\mathrm{mae}(S)$ unique, but this is not necessary for $S = A^+_\Sigma$: since any of

Definition 10.34 *For any surface $S \subset \Sigma$ (in the above sense), a* **minimal area enclosure** $\mathrm{mae}(S)$ *is a surface such that* $\mathrm{mae}(S) \supset S$, *and* $\mathrm{Area}(\mathrm{mae}(S)) \leq \mathrm{Area}(S')$ *for all surfaces* $S' \supset S$.

Thus we replace A^+_Σ by $\mathrm{mae}(A^+_\Sigma)$, which exists and, being an extremizing surface, saturates the inequality $\theta^+ \leq 0$. Thus $\mathrm{mae}(A^+_\Sigma)$ is an MOTS. Taking $S = A^+_\Sigma$ and $S' = H^+_E \cap \Sigma$ we see that

$$\mathrm{Area}(\mathrm{mae}(A^+_\Sigma)) \leq \mathrm{Area}(H^+_E \cap \Sigma), \tag{10.156}$$

as desired. Hence we may take $\tilde{A}_\Sigma = \mathrm{Area}(\mathrm{mae}(A^+_\Sigma))$ in our earlier discussion. Thus we put:

Definition 10.35 *For any asymptotically flat initial data set $(\Sigma, \tilde{g}, \tilde{k})$ satisfying (8.110), with associated asymptotic mass (10.151) and apparent horizon A^+_Σ, the* **Penrose inequality** *is*

$$\mathrm{Area}(\mathrm{mae}(A^+_\Sigma)) \leq 16\pi m^2_\Sigma. \tag{10.157}$$

In the static case, i.e. $\tilde{k} = 0$, the following simplifications take place (cf. Definition 8.4);

1. The initial data set $(\Sigma, \tilde{g}, \tilde{k})$ becomes an asymptotically flat Riemannian manifold $(\Sigma, \tilde{g})$;
2. The assumption (8.110) on the initial data becomes $\tilde{R} \geq 0$, where $\tilde{R}$ is the Ricci scalar;
3. The apparent horizon A^+_Σ becomes the ***outermost minimal surface*** A_Σ in Σ, i.e., the (unique) minimal surface such that no other minimal surface in Σ properly encloses A_Σ.[622]
4. In computing the area there is no need for the minimal area enclosure.

One can see this for the spatial Schwarzschild metric, provided one uses the radial coordinate ρ instead of r, since the metric (10.117) is singular at the place of interest $r = 2m$, see (10.122). By spherical symmetry it is enough to consider radial perturbations: the area $4\pi r(\rho)^2$ is minimized iff $\rho = m/2$, which corresponds to $r = 2m$ and hence recovers the apparent = event horizon.[623]

Theorem 10.36 *Any complete asymptotically flat 3d Riemann manifold $(\Sigma, \tilde{g})$ with $\tilde{R} \geq 0$, with asymptotic mass $m_\Sigma = \Pi^0$ defined by (8.103), satisfies the* **Riemannian Penrose inequality**

$$\mathrm{Area}(A_\Sigma) \leq 16\pi m^2_\Sigma, \tag{10.158}$$

where A_Σ is the unique outermost minimal surface in Σ (assumed to have one end only).[624] *Furthermore ("rigidity"), equality in (10.158) holds iff the region outside A_Σ is isometric to the part $r > 2m$ of the Schwarzschild space $(\Sigma'_S, \tilde{g}_S)$ defined in and above (10.117).*

We have to refer to the literature for a proof of this.[625] Meanwhile, the general case in Definition 10.35 seems out of reach (it has been proved only for spherically symmetric space-times).

---

its minimal area enclosures is an MOTS, by definition $\mathrm{mae}(A^+_\Sigma)$ encloses, and hence must coincide with, any rival. Technically, $\mathrm{mae}(A^+_\Sigma)$ is not just an MOTS but an *outermost* MOTS, which is unique if it exists. Even its possible non-uniqueness would not affect the Penrose inequality (10.157), since any two candidates have the same area.

[622] Equivalently, $A_\Sigma = \partial(\cup\{U \in \mathscr{O}(\Sigma) \mid \partial U \text{ is a minimal surface}\})$, where "surface" is meant as explained above (10.153). See Theorem 4.7 in Lee (2019) for existence and uniqueness of A_Σ. It can be shown that provided $\tilde{R} \geq 0$, outermost minimal surfaces are two-spheres (Meeks & Yau, 1980). Note that $\dim(\Sigma) = 3$ throughout this section, but *mutatis mutandis* result like this are valid up to $\dim(\Sigma) < 8$ (at $n \geq 8$ smoothness of A_Σ turns out to be lost).

[623] Note that the spatial Schwarzschild metric $\tilde{g}_S$ is complete on the full space $\Sigma_S = \mathbb{R}^3 \backslash \{0\}$ on which it is defined.

[624] This assumption can be dropped by taking the outermost minimal surface with respect to some given end. See Lee (2019), Conjecture 4.12, which also generalizes the inequality to arbitrary dimension.

[625] See Huisken & Ilmanen (1997, 2001) and Bray (2001), as well as the reviews cited in footnote 614.

10.12 Epilogue: The laws of black hole thermodynamics

Around 1970, it was noted by various people that the following dictionary made some sense:[626]

Thermodynamics	Black Holes
equilibrium state σ	*stationary metric* g
temperature T	*surface gravity* κ
entropy S	*horizon surface area* A
energy E	*asymptotic mass* m
other conserved quantities	*(Komar) asymptotic quantities*

The basis for this analogy lies in the following three ***laws of black hole thermodynamics***:[627]

Zeroth law: The surface gravity is constant on each connected component of the event horizon.

First law: For simplicity taking just one conserved quantity into account, viz. angular momentum J,

$$\frac{\kappa}{8\pi}\delta A = \delta m - \Omega_H \delta J, \tag{10.159}$$

where Ω_H is a constant at the event horizon (playing the role of a chemical potential).

Second law: Hawking's ***area law***,[628] i.e.

$$\delta A \geq 0. \tag{10.160}$$

These were initially seen as laws of black hole *mechanics*. Despite powerful arguments by Bekenstein, the possibility of a true thermodynamic underpinning was even explicitly denied:[629]

[626] See Thorne (1994) and Weinstein (2021) for some of the history of black hole thermodynamics; pioneering papers include Christodoulou (1970), Christodoulou & Ruffini (1971), Penrose & Floyd (1971), Hawking (1972), Bekenstein (1972, 1973, 1974), and Bardeen, Carter & Hawking (1973), which stated all four laws.

[627] The zeroth law of classical thermodynamics states that (thermal) equilibrium is an equivalence relation, which is what allows the introduction of temperature T in the first place, and then implies that T is constant in thermal equilibrium. The first law (or, if seen as "conservation of energy", a consequence thereof) is $T\delta S = \delta E + \sum_i \mu_i \delta Q_i$, where the Q_i are the relevant conserved quantities and the μ_i their (generalized) chemical potentials (for example, we count volume V as one such Q_i, with $\mu_i = p$). The second law is $\delta S \geq 0$, one of the great mysteries of physics. We omit the third law, which states that κ cannot be brought to zero by a 'finite sequence of operations' (Bardeen, Carter & Hawking, 1973) or 'within a finite advanced time' (Israel, 1986). This idea is physically ambiguous if not disputable and also lacks a clear connection with the usual version of the third law of thermodynamics, to the effect that the entropy is zero at zero temperature (which would even be violated by extremal black holes).

[628] Continuing footnote 485 on the history of the definition (10.79) of the "absolute" event horizon, i.e. $H_E^+ := \partial I^-(\mathscr{I}^+)$, left as something between Hawking and Penrose: in Seife (2021, p. 478 of e-book) Penrose recalls a telephone conversation he had with Hawking in 1970 in which they discussed the area law including the crucial role of the definition of the horizon, which Hawking proposed to Penrose but followed this with: 'it was your idea' Penrose adds: 'I don't know what he thought. Maybe he thought I had the idea but didn't quite have it. It's not clear. I don't know what the story was, really. I never wanted to bring it up. Because it was a big thing for him.'

[629] We read in Seife (2021), chapter 13, that the word "mechanics" in the title 'The four laws of black hole mechanics' of Bardeen, Carter, & Hawking (1973) was a deliberate provocation against Bekenstein (whose name they even misspelled as Beckenstein), who first proposed that the analogy between the pertinent laws for black holes and the laws of (ordinary) thermodynamics was more than a purely formal one, and hence has physical content. Especially Hawking, who had discovered his singularity theorem and the second law, among other things, and (perhaps with hindsight) was well on his way to fame and fortune, initially responded quite harshly to Bekenstein, who at the time was just a PhD student (though not an entirely powerless one, as his supervisor, Wheeler, who at the time had a significant if not controlling influence on the Western GR community, took his side).

> It should however be emphasized that $\kappa/8\pi$ and A are distinct from the temperature and entropy of the black hole. In fact the effective temperature of a black hole is absolute zero. (Bardeen, Carter, & Hawking, 1973, p. 168)

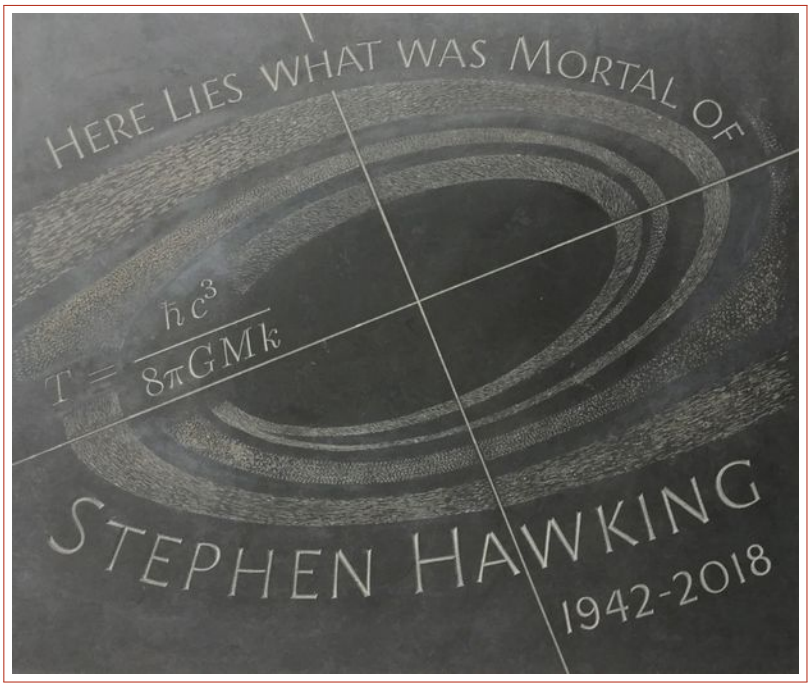
However, within a year Hawking made a remarkable U-turn, which changed physics. On the basis of a calculation in quantum field theory in curved space-time involving pair creation near the horizon, he predicted the radiation now named after him, which turns a black hole into a black body and (allegedly) shows that the laws of black holes *mechanics* are genuinely laws of black hole *thermodynamics*. Hawking's calculation also allowed the explicit identifications

$$S = \frac{k_B c^3}{4G\hbar} A \equiv A/4; \tag{10.161}$$

$$k_B T = \frac{\hbar}{2\pi c} \kappa \equiv \kappa/2\pi, \tag{10.162}$$

called the ***Bekenstein–Hawking entropy*** and the ***Hawking temperature***, respectively.[630] For example, for a Schwarzschild black hole, where $\kappa = c^4/4mG \equiv 1/4m$, for the latter we obtain

$$T = \frac{\hbar c^3}{8\pi G m k_B} \; (\approx 6 \cdot 10^{-8} \; ^\circ K). \tag{10.163}$$



Tombstone of Stephen Hawking's grave in Westminster's Abbey, containing his ashes. At his 60th birthday in 2002, Hawking requested equation (10.163) *to be engraved on his tombstone.*

Note the spectacular combination of fundamental constants,[631] hidden by the use of "natural" units $G = \hbar = c = k_B = 1$ on the right in (10.161) - (10.162). For a Schwarzschild black hole one has $A = 4\pi r_S^2$ with $r_S = 2Gm/c^2 \equiv 2m$, and hence its dimensionless entropy equals

$$S/k_B = 4\pi G m^2/\hbar c \; (\approx 10^{77}). \tag{10.164}$$

[630] Bekenstein (1973) gave a similar formula for the temperature of a Kerr black hole, whose Schwarzschild limit differs from Hawking's by a multiplicative constant. Hawking's formula first appeared, in the form (10.162), in Hawking (1974). Thus also the temperature T is sometimes named after both Bekenstein and Hawking.

[631] See e.g. `https://www.vttoth.com/CMS/physics-notes/311-hawking-radiation-calculator`.

We will not discuss the microscopic justification of black hole thermodynamics, since unlike black hole thermodynamics itself, all attempts to underpin it use quantum (field) theory or string theory etc. and hence blast the framework of classical GR.[632] Within the classical setting, it cannot be overemphasized how closely this topic is related to black hole uniqueness theorems (cf. the previous two sections), since the key to both classical thermodynamics and black hole thermodynamics lies in the fact that just a few "emergent" parameters control the situation.

Black hole thermodynamics comes with all the problems of classical thermodynamics, starting with the question what the symbol "δ" in the first and second laws is supposed to mean. Amidst the huge literature on thermodynamics and its foundations, mathematical physicists typically prefer the axiomatization of Lieb and Yngvason.[633] This is restricted to equilibrium states and transitions between these through processes that fall under the following heading:

> **Adiabatic accessibility:** A state Y is adiabatically accessible from a state X (...) if it is possible to change the state from X to Y by means of an interaction with some device (which may consist of mechanical and electrical parts as well as auxiliary thermodynamic systems) and a weight, in such a way that the device returns to its initial state at the end of the process. (Lieb & Yngvason, 1999, p. 17)

In contrast, the weight may have risen or fallen (in a gravitational field). This incorporates the original thermodynamic idea of a cycle, but avoids the equally traditional but mysterious concept of "heat", which in any case is problematic for black holes. The definition of adiabatic accessibility also includes what in black holes thermodynamics is called the "physical process" interpretation, in which one studies what happens if things are thrown into a black hole.[634]

Thus the "δ" in the laws of black hole thermodynamics should, in principle, refer to changes in selected properties of a black hole metric (viz. its asymptotic mass, angular momentum, possibly charge, and spatial horizon area) if, due to some intervention, the metric evolves from one stationary value to another. Unfortunately, this only seems to apply to the first law (which is predicated on the zeroth). A typical application of the second law, mentioned from the beginning by Hawking and others, is the merger of two black holes into one, whose area, then, is greater than or equal to the sum of the areas of the original constituents.[635] Since multi-black hole metrics are typically unstable (except for the charged Majumdar–Papapetrou metric studied in §10.9) it seems impossible to see this merger as an adiabatic evolution of the said type.[636]

This, and in fact the whole theory, suggests that each of the laws of black hole thermodynamics is valid in its own unique setting, and that there is not, as far as we know, a single setting–formalized as a set of mathematical assumptions–in which all laws are valid. This nature of black hole thermodynamics as patchwork will be reflected by the following discussion, which starts out historically and then, in vain, tries to converge to a more systematic presentation.

---

[632] Surveys of black hole thermodynamics include Wald (1994, 2001), Jacobson (1996), Compère (2006), Bravetti (2014), Carlip (2014), Curiel (2014b), Dougherty & Callender (2016), Wall (2018), and Wallace (2018, 2019).

[633] See Lieb & Yngvason (1999), which (as they note) was partly inspired by Planck (1926) and Giles (1964). For the connection with (classical) statistical mechanics see also Martin-Löf (1979) and Uffink (2007).

[634] See Wald (1994) and Gao & Wald (2001). This interpretation complements Bardeen, Carter & Hawking (1973), who study (asymptotic) parameter changes of unidentified origin for 'two slightly different stationary axisymmetric black hole solutions', clearly inspired (via the uniqueness theorems) by the Kerr–Newman metric. Their proofs suggest that these changes always pass through other such solutions, whereas the "physical process" interpretation makes no such assumption, as long as eventually a new equilibrium state = stationary metric is reached.

[635] Hawking (1972) shows that the opposite process, i.e. a bifurcation of one black hole into various black holes, cannot happen (even if it were compatible with the area law). See also Hawking & Ellis (1973), Proposition 9.2.5.

[636] Black hole mergers are the source of the gravitational waves detected on earth (Castelvecchi, 2020).

With hindsight, black hole thermodynamics started with the ***Penrose process*** discussed in §9.6. In the spirit of "$E = mc^2$" we may opportunistically rewrite the inequality (9.153) as

$$\delta m - \Omega_+ \delta J > 0, \tag{10.165}$$

where m is the mass and $J = am$ is the angular momentum of the (Kerr) black hole, cf. (9.120). Eq. (10.146) now gives ***Smarr's formula*** as well as, with more work, its variational form

$$\frac{\kappa_+}{4\pi} A_K = m - 2\Omega_+ J; \qquad \frac{\kappa_+}{4\pi} \delta A_K = 2(\delta m - \Omega_+ \delta J), \tag{10.166}$$

where the surface gravity κ_+ at the outer (event) horizon is given by (10.110). To derive the second part, one should first express κ_+ and Ω_+ in the independent variables m and J via (9.120), (9.143), (9.117), and (10.110), and then put them back at the end of the calculation.

Clearly, eq. (10.166) is a special case of the first law of black hole thermodynamics, and if we combine it with (10.165) we also obtain an example of the second law (10.160), i.e. the Hawking area theorem, at least in the special case that the change of area is caused by the Penrose process.

A more general argument for (10.160) is obtained by turning Proposition 6.14 on its head.[637] Originally intended to prove that the assumption $\theta(x) < 0$ leads to cusps or focal points in the null hypersurface C defined by (6.61), if we instead assume that C is smooth and that its null generators are future complete, then $\theta(x) < 0$ leads to a contradiction, so that under the stated assumptions we must have $\theta(x) \geq 0$. We apply this to the case where C is some component of the event horizon H_E^+ of a black hole region as defined in (10.79), which is not necessarily stationary and may even consist of various components, that is, of various black holes (which may merge). The structure of these components is described by Proposition 10.16.1, which shows that C is ruled by future inextendible lightlike geodesics γ. In order to apply Proposition 6.14 (contrapositively) we now need to argue that each component C is smooth (which is not automatic, since from Proposition 6.18 we merely know it is locally Lipschitz). This should follow from additional assumptions, such as some form of weak cosmic censorship that prevents the inextendibility of the lightlike geodesics that rule C to be caused by incompleteness.[638]

It is also assumed that one can foliate at least the relevant part of space-time by partial (i.e. "wannabe") Cauchy surfaces Σ_t, which intersect each component C of H_E^+ in a two-sphere S_t (cf. Proposition 10.29). In other words, each S_t is a spatial cross-section of some component of the event horizon, whose area $\mathrm{Area}(S_t)$ is defined by (6.75). As already remarked, smoothness of C then enforces $\theta \geq 0$. Under the assumptions of Proposition 6.14, notably the null curvature condition, eq. (6.76) then implies that each area $\mathrm{Area}(S_t)$ and hence also their sum A_t, i.e. the total area of $H_E^+ \cap \Sigma_t$, can only increase with time (or stay the same). We may write this as

$$\Sigma_{t_1} \subset J^-(A_{t_2}) \qquad \Rightarrow \qquad \mathrm{Area}(H_E^+ \cap \Sigma_{t_1}) \leq \mathrm{Area}(H_E^+ \cap \Sigma_{t_2}). \tag{10.167}$$

This is a precise version of the second law, from which the problematic "δA" has been removed.

[637] See Hawking (1972) and Hawking & Ellis (1973), §9.2. Various sets of assumptions are known not merely for a rigorous *proof* of the area theorem, but even for its *formulation*, since the lack of smoothness of the horizon (which *a priori* is only locally Lipschitz, cf. Proposition 6.18) requires conditions making the area (6.75) *well defined*. See Chruściel *et al.* (2001) for a detailed analysis of various assumptions, including the corresponding proofs. The simplest–though by no means the weakest–assumption is global hyperbolicity of the conformal completion $(\hat{M}, \hat{g})$ of the given asymptotically flat space-time (M, g), see Definition 10.1, along with future completeness of the null generators of the horizon H_E^+ and validity of the null curvature condition (as in Theorem 6.15) on $I^-(\mathscr{J}^+)$.

[638] In Hawking's proof this form was *strong asymptotic predictability*, which roughly speaking means that $\overline{I^-(\mathscr{J}^+)}$ is contained in a globally hyperbolic region (Hawking & Ellis, 1973, p. 313). See also Wall (2009), §1.2.3.

Turning to the (later) zeroth law, the key observation was that under various assumptions one can sharpen Proposition 10.21 to constancy of the surface gravity κ on the entire Killing horizon. The relevance of this result follows from Hawking's rigidity theorem in §10.10, which makes the event horizon a Killing horizon. The simplest such result is as follows.[639]

Proposition 10.37 *The surface gravity κ is constant and nonzero on each component of a bifurcate Killing horizon H_K, and differs at most by a sign on different components thereof.*

Proof. As in the proof of Proposition 10.21, from (10.106) and (10.104) we obtain

$$\mathscr{L}_{e_I}\kappa^2 = e_I^\mu \partial_\mu \kappa^2 = -R^\alpha_{\nu\mu\beta} e_I^\mu X^\beta \cdot \nabla^\nu X_\alpha, \tag{10.168}$$

where $I = 1,2$ and the spacelike unit vectors e_I form an orthogonal basis of the orthogonal complement of X at each $T_x H_K$, $x \in H_K$ (cf. Lemma 4.16). Since $X = 0$ at the bifurcation surface S, it follows from (10.168) that κ^2 is constant on S. Since different lightlike geodesics ruling H_K emanate from different points of S, eq. (10.100) implies that κ^2 is constant on H_K altogether.

To show that $\kappa \neq 0$, we note, proving by contradiction, that $\kappa = 0$ and (10.106) imply $\nabla_\mu X_\nu = 0$ on S, because the spacelike contractions in (10.106) vanish by themselves and hence the total expression is negative semidefinite.[640] Hence $\nabla_\mu X_\nu$ and X_ν both vanish on S, but this implies that X is identically zero (so that it could not be lightlike, see §5.3).[641] □

The zeroth law exemplifies the fact that the laws of black hole thermodynamics may be derived under various inequivalent assumptions, since the original version was as follows:[642]

Proposition 10.38 *If the Einstein equations (7.1) and the dominant energy condition (7.65) hold, then the surface gravity κ is constant on any (necessarily connected) Killing horizon H_K.*

Proof. If the null generators of H_K have tangent vectors L, then by Lemma 4.16 we have on H_K:

$$X = f \cdot L, \tag{10.169}$$

where f is some function defined on H_K. From (10.99) and $\nabla_L L = 0$, we find $\kappa = Lf$. Assuming that H_K is sufficiently smooth, the Frobenius condition (8.94) for (null) hypersurface orthogonality of the Killing vector field X and (6.88) then give $k_{\mu\nu} = 0$ on H_K, so that also $\theta = 0$ and $\sigma_{\mu\nu} = 0$ on H_K. The null Raychaudhuri equation (6.98) then gives

$$R_{\mu\nu} X^\mu X^\nu = 0. \tag{10.170}$$

This also follows by noting that the area $\mathrm{Area}(S_t)$ of a stationary black hole must be independent of t, so that (6.76) gives $\theta = 0$ and hence $\dot\theta = 0$, after which (6.98) again yields (10.170).

[639] The assumption of a bifurcate Killing horizon is not very heavy; Rácz & Wald (1996) give arguments 'supporting the view that any spacetime representing the asymptotic final state of a black hole formed by gravitational collapse may be assumed to posses a bifurcate Killing horizon or Killing horizon with vanishing surface gravity' (the latter occurs in extremal Kerr and Reissner–Nordström black holes, whose existence astrophysicists deny).

[640] On S we have $\nabla_{e_I} X^\mu = e_I^\alpha(\partial_\alpha X^\mu + \Gamma^\mu_{I\beta} X^\beta) = 0 + 0 = 0$, as X vanishes on S and $e_I^\alpha \partial_\alpha = e_I$ is tangent to S.

[641] Any isometry ψ of M is determined at least locally (i.e. in a convex nbhd of x) by its tangent map ψ'_x at some fixed $x \in M$: to find $\psi(y)$, take the unique geodesic γ from x to y, so that $y = \exp_x(Y)$ for some $Y \in T_x M$, and if ψ is an isometry, then $\psi(\exp_x(Y)) = \exp_x(\psi'_x(Y))$. Infinitesimally, this implies that any Killing vector field X is determined by $X(x)$ and $\nabla_\mu X_\nu(x)$. See also Chruściel (2020), Proposition 4.3.10, for a direct proof of this.

[642] See Bardeen, Carter & Hawking (1973), §2, as well as Wald (1984), §12.5, and Chruściel (2020), §4.3.4.

Eq. (10.170) holds on H_K, where, using the Einstein equations (7.1), it implies that $T_{\mu\nu}X^\mu X^\nu = 0$. Hence the vector $T(X)$, with components $T(X)^\mu := T^\mu_\nu X^\nu$, is orthogonal to X, since

$$g(T(X),X) = T_{\mu\nu}X^\mu X^\nu = 0. \tag{10.171}$$

Therefore, by Lemma 4.16 this vector is either spacelike, or lightlike and hence proportional to X, or zero. On the other hand, since X is lightlike and hence causal, the dominant energy condition (7.65) forces $T(X)$ to be causal or zero. This excludes the possibility that $T(X)$ is spacelike and hence it must be null, all of this on H_K only. Again invoking Lemma 4.16, we conclude that $T(X)$ must be proportional to X. Using (7.1) in the opposite direction gives

$$X^\flat \wedge R(X)^\flat = X^\flat \wedge (T(X) - \tfrac{1}{2}T\cdot X)^\flat = 0. \tag{10.172}$$

The final step in the proof is the following equality, which as usual in this proof is valid on H_K:

$$X^\flat \wedge R(X)^\flat = d\kappa \wedge X. \tag{10.173}$$

To prove this, we note that since X is a Killing vector field, eq. (10.99) is equivalent to

$$X^\mu \nabla_\nu X_\mu = -\kappa X_\nu, \tag{10.174}$$

cf. (3.74). Eq. (10.173) follows by applying the antisymmetrized expression $X_{[\rho}\nabla_{\sigma]}$ to both (10.174) and (8.94) and carrying out some lengthy but straightforward rearrangements. Eqs. (10.172) and (10.173) yield $d\kappa \wedge X = 0$ on H_K, which forces $d\kappa = 0$ on H_K. □

We now return to the first law of black hole thermodynamics (10.159), which "morally" reads

$$T\delta S = \delta E - \Omega_H \delta J. \tag{10.175}$$

where we identify $m = E$ and regard the constant Ω_H as a generalized chemical potential. The mass/energy m/E and the angular momentum J of a black hole are defined by the Komar formulae (9.118), see below. Referring to the Penrose process discussed above for at least an example of the "physical process" interpretation of (10.175), we now give a derivation based on the idea that the δ's indicate that one gently moves the metric from one stationary value to another through intermediate stationary metrics. This derivation is based on the Hamiltonian formalism of GR.[643] Like Proposition 10.37, the argument requires a bifurcate Killing horizon, but as argued after (10.108) and in footnote 639 this is not really a very strong assumption.

More seriously, the proof relies on Hawking's rigidity theorem (i.e. Theorem 10.28 in §10.10), to the effect that the event horizon is a Killing horizon for a Killing vector field

$$X = \partial_t + \Omega_H \partial_\varphi, \tag{10.176}$$

which generalizes the one for the Kerr metric, cf. (9.149). This means that this particular derivation of the first law also requires the assumptions of the rigidity theorem, which include,

[643] We follow Sudarsky & Wald (1992). The existence of the spacelike surface Σ used in the proof was proved by Chruściel & Wald (1994b). The equivalence between the Hamiltonian (ADM) versions of the asymptotic mass and angular momentum (which only holds in stationary asymptotically flat space-times) is discussed in Jaramillo & Gourgoulhon (2009); see also Poisson (2004), §4.3 and Gourgoulhon (2012), chapter 8. A more elegant derivation can be given from the Lagrangian formalism and its associated covariant Noether charges, originally developed by Kijowski & Tulczyjew (1979). See Wald (1993), Iyer & Wald (1994), and Jacobson, Kang, & Myers (1994). See Gao & Wald (2001) and Poisson (2004), §5.5.3 and §5.5.4 for "physical process" derivations of the first law.

for example, a version of weak cosmic censorship and imply, among other things, that (M,g) is axisymmetric, with Killing vector field ∂_φ (the special case $X=\partial_t$, i.e. the Killing vector field defining stationarity, is just $\Omega_H=0$). In general, Ω_H is some constant (interpreted as the angular velocity of the black hole) chosen so that indeed $g(X,X)=0$ at the event horizon, see §10.10.

Recall eqs. (8.214) - (8.222) from the Hamiltonian approach to GR, in which we take Σ to be a spacelike surface whose boundary at one end is the given bifurcation surface $\mathscr{S}$, and at the other end is a two-sphere S^2_r as in (8.126), where we eventually let $r\to\infty$; for simplicity we just write this boundary component as S^2_∞. Using manipulations similar to those in the derivation of (7.44) but now in one dimension lower, we may rewrite the boundary Hamiltonian (8.220) as

$$H_B(\Sigma)=\int_{S^2_\infty\cup\mathscr{S}} d^2\sigma^i\,(L(\tilde\nabla^j\tilde g_{ij}-\tilde\nabla_i\tilde g^j_j)+2S^j\check\pi_{ij})=:H_B(S^2_\infty)+H_B(\mathscr{S}), \tag{10.177}$$

where L is the lapse and $S=S^i\partial_i$ is the shift, so far arbitrary. The trick is to choose these as

$$LN+S=X, \tag{10.178}$$

cf. (8.5), where N is the future-pointing normal to Σ as usual. This has the effect that

$$\frac{\partial\tilde g_{ij}}{\partial t}=0; \qquad\qquad \frac{\partial\tilde\pi^{ij}}{\partial_t}=0, \tag{10.179}$$

since the time evolution generated by (10.178), i.e. the flow of X, consists of isometries.

Now consider variations of $H_G(\Sigma)$, see (8.218) and (8.221), induced by one-parameter families (homotopies) $\tilde g^s_{ij}$ and $\tilde\pi^{ij}_s$ that satisfy the constraints (8.225). Then the variations

$$\delta\tilde g_{ij}:=\frac{d\tilde g^s_{ij}}{ds}(s=0); \qquad\qquad \delta\tilde\pi^{ij}:=\frac{d\tilde\pi^{ij}_s}{ds}(s=0) \tag{10.180}$$

satisfy the linearized constraint equations, i.e. $C_0'(\delta\tilde g,\delta\tilde\pi):=dC_0(\tilde g^s,\tilde\pi_s)_{|s=0}=0$, etc. Then

$$\delta H_G(\Sigma)=0 \tag{10.181}$$

by (8.221). On the other hand, $\delta H_G(\Sigma)$ consists of a bulk term giving the equations of motion (8.227) and (8.228) and a boundary term. By (10.179) the former vanishes, and so we must have

$$\delta H_B(\Sigma)=0. \tag{10.182}$$

The Killing vector field X in Theorem 10.28 is such that at (spatial) infinity we have $N\to 1$ and $S\to\partial/\partial_\varphi$, and hence by definition of these quantities the integral over S^2_∞ in (10.177) gives

$$\delta H_B(S^2_\infty)=16\pi(\delta E-\Omega_H\delta J), \tag{10.183}$$

cf. (8.126).[644] To compute the integral over the bifurcation surface $\mathscr{S}$, we perform a partial integration and realize that by definition $X=0$ and hence $L=S^i=0$ on $\mathscr{S}$. This simply gives

$$\delta H_B(\mathscr{S})=-\int_{\mathscr{S}} d^2\sigma_i\,(\partial_j L)\cdot(\tilde g^{ij}\tilde g^{kl}-\tilde g^{ik}g^{jl})\delta\tilde g_{kl}, \tag{10.184}$$

[644] The factor 16π comes from the fact that the Einstein–Hilbert action (7.2) should really be $(c^4/16\pi G)\int R$.

where $d^2\sigma_i = d^2\sigma \cdot n^i$, with n the *inward* normal to $\mathscr{S}$ within Σ (see footnote 646), and

$$d^2\sigma = d^2z\sqrt{\det(\check{g}(z))}, \tag{10.185}$$

as in (8.220). Since $X = 0$ and hence $L = 0$ on $\mathscr{S}$, cf. (10.178), we have $\partial_j L = n_j \nabla_n L$, so that

$$\delta H_B(\mathscr{S}) = -\int_{\mathscr{S}} d^2\sigma\,(\nabla_n L)\cdot(\tilde{g}^{kl} - n^k n^l)\delta\tilde{g}_{kl} = -2\delta\int_{\mathscr{S}} d^2\sigma\,\nabla_n L, \tag{10.186}$$

where the second equality follows as in (7.34), with the additional remark that $(\tilde{g}^{kl} - n^k n^l)\tilde{g}_{kl}$ is the "covariant" expression for the metric $\check{g}$ on $\mathscr{S}$, as in (7.38) but in one dimension higher (throughout this derivation, δ acts only on $\tilde{g}$ and $\tilde{\pi}$, not on L and S). We now use the identity

$$2\int_S d^2\sigma\,(\nabla_n L - \tilde{k}_{ij}n^i S^j) = -\int_S d\sigma_{\mu\nu}\nabla^\mu X^\nu, \tag{10.187}$$

which is valid on any 2-surface S, and relates H_B to the Komar definition of a conserved quantity defined through any Killing vector field X, cf. (9.118).[645] This time the surface element is

$$d\sigma_{\mu\nu} = (X_\mu \underline{X}_\nu - X_\nu \underline{X}_\mu)\,d^2\sigma, \tag{10.188}$$

where X is seen as a lightlike vector on the Killing horizon and hence is complemented by another lightlike vector $\underline{X}$ orthogonal to $\mathscr{S}$, cf. §6.3, which unlike (6.58) is normalized such that

$$g(X,\underline{X}) = X^\mu \underline{X}_\mu = -1. \tag{10.189}$$

Furthermore, $\tilde{k}_{ij}$ is given by (8.24), but in view of (8.210) we can ignore the term $\tilde{k}_{ij}n^i S^j$ since $S^j = 0$ on the bifurcation surface $S = \mathscr{S}$. Using (10.188), (10.99), and (10.189) we obtain

$$\int_S d\sigma_{\mu\nu}\nabla^\mu X^\nu = -\kappa A, \tag{10.190}$$

as the second term in (10.188) gives zero because $X^\mu X_\mu = 0$ implies $X^\nu \nabla^\mu X_\nu = 0$, and we have taken κ out of the integral since it is constant by the zeroth law of black hole thermodynamics, i.e. Proposition 10.37.[646] Since δ in (10.186) only acted on the surface element, we find

$$\delta H_B(\mathscr{S}) = -2\kappa\delta A. \tag{10.191}$$

Eqs. (10.182), (10.183), (10.186), and (10.191) then recover (10.159), which using the identifications (10.161) and (10.162) is the first law of black hole thermodynamics (10.175).

The situation covered by this proof of the first law is quite different from its original Penrose process context, in which particles were thrown into a Kerr black hole. Perhaps because it is the frontier of fundamental physics, black hole thermodynamics is also a gallimaufry of ideas.

[645] See e.g. Jaramillo & Gourgoulhon (2009), eq. (16). This is a fairly easy exercise. See also footnote 646.

[646] See Poisson (2004), §5.5.3. This shows that $\nabla_n L = \kappa$, as stated without proof in Sudarsky & Wald (1992). The sign of κ depends on the branch of the bifurcate Killing horizon, and hence on the sign of the normal n to $\mathscr{S}$ within Σ. It can be checked on the example (10.93) in $4d$ that one needs the *inward* normal, which gives the minus sign in (10.184), since Stokes's theorem (i.e. partial integration) uses the *outward* normal. In this example, the bifurcation surface is $x = t = 0$, i.e. the y-z plane, and $L = x$. The correct normal for the future horizon $x = t$, where $\kappa = 1$, is $n = \partial/\partial x$, which is directed *inward*, and indeed we duly obtain $\nabla_n L = \partial x/\partial x = 1 = \kappa$.

A Lie groups, Lie algebras, and constant curvature

This appendix contains material supporting §4.4 on spaces of constant curvature, but is also interesting elsewhere. Its content underpins much of mathematics and mathematical physics.[647]

A.1 Lie groups

We only need real *linear* ***Lie groups***, which are closed subgroups of $GL_n(\mathbb{R})$, i.e. the group of real invertible $n\times n$ matrices, with group multiplication simply given by matrix multiplication.[648] For example, $SO(3)$ is the subgroup of $GL_3(\mathbb{R})$ consisting of matrices R that satisfy

$$R^T R = 1_3; \tag{A.1}$$
$$\det(R) = 1. \tag{A.2}$$

More generally, for some given $\Gamma \in GL_n(\mathbb{R})$, the matrices $\gamma \in GL_n(\mathbb{R})$ that for all x,y satisfy

$$\langle \gamma x, \Gamma \gamma y\rangle = \langle x, \Gamma y\rangle, \tag{A.3}$$

or, in other words, leave the bilinear form $\langle x,y\rangle_\Gamma = \langle x,\Gamma y\rangle$ invariant (where $\langle\cdot,\cdot\rangle$) is the usual inner product on $\mathbb{R}^n$), form a linear Lie group G_Γ. In other words,

$$G_\Gamma = \{\gamma \in GL_n(\mathbb{R}) \mid \gamma^T \Gamma \gamma = \Gamma\}. \tag{A.4}$$

For $n=3$ and $\Gamma = 1_3$ we obtain $G_\Gamma = O(3)$, which has has two components: the one containing the identity is $SO(3) \equiv O(3)_+$, singled out by $\det(R)=1$, whereas the other component $O(3)_-$ consists of those elements $R\in O(3)$ with $\det(R)=-1$. Note that $SO(3)$ is connected but not *simply* connected. Furthermore, $O(3)$ and $SO(3)$ are *compact* in the topology inherited from $M_3(\mathbb{R}) \cong \mathbb{R}^9$: this follows from the following parametrization of $SO(3)$, with $\alpha,\beta,\gamma\in[0,2\pi]$:

$$R^z_\gamma = \begin{pmatrix} \cos\gamma & -\sin\gamma & 0 \\ \sin\gamma & \cos\gamma & 0 \\ 0 & 0 & 1 \end{pmatrix}, R^y_\beta = \begin{pmatrix} \cos\beta & 0 & -\sin\beta \\ 0 & 1 & 0 \\ \sin\beta & 0 & \cos\beta \end{pmatrix}, R^x_\alpha = \begin{pmatrix} 1 & & 0 \\ 0 & \cos\alpha & -\sin\alpha \\ 0 & \sin\alpha & \cos\alpha \end{pmatrix}.$$

Staying in $n=3$ for the moment, instead of $\Gamma = 1_3$ we may take $\Gamma = \mathrm{diag}(-1,1,1)$. Then $G_\Gamma \equiv O(1,2)$ is called the ***Lorentz group*** (in space-time dimension 3). It has *four* components, singled out by the four combinations of the two independent conditions

$$\det(\lambda) = \pm 1; \qquad\qquad \pm\lambda_{00} > 0; \tag{A.5}$$

for an indefinite matrix Γ like this it is customary to label the entries λ_{ij} by $i,j=0,1,2$ instead of $1,2,3$. In particular, the identity component $O(1,2)_0$ satisfies $\det(\lambda)=1$ and $\lambda_{00}>0$.[649] Consequently, even the subgroup $SO(1,2) = \{\lambda\in O(1,2) \mid \det(\lambda)=1\}$ has *two* components.

[647] References for this appendix are Helgason (1978), O'Neill (1983), Vinberg (1993), and Wolf (2011).

[648] Such Lie groups are not necessarily closed in $M_n(\mathbb{R})$, since invertibility of matrices is an open condition (we call a condition *open* if its solution set is open, and *closed* if its solution set is closed). For example, the sequence $g_n = 1_n/n$ in $GL_n(\mathbb{R})$ converges to zero, so the limit is not in $GL_n(\mathbb{R})$. The topology used may either be the usual one on $\mathbb{R}^{n^2}$ or the matrix norm topology; these are equivalent.

[649] This follows from the fact that any matrix $\lambda\in O(1,2)$ satisfies $\lambda_{00}^2 - \lambda_{10}^2 - \lambda_{20}^2 = 1$, so that $|\lambda_{00}|\geq 1$, and from the fact that $\mathrm{sgn}(\lambda_{00})$ and $\det(\lambda)$ are continuous functions on $O(1,2)$.

Another important difference with $SO(3)$ is that $SO(1,2)$ is *non-compact*. This follows, for example, from the following parametrization of $O(1,2)_0$, where $\alpha \in [0,2\pi]$ and $\beta,\gamma \in \mathbb{R}$:

$$B^x_\gamma = \begin{pmatrix} \cosh\gamma & \sinh\gamma & 0 \\ \sinh\gamma & \cosh\gamma & 0 \\ 0 & 0 & 1 \end{pmatrix}, B^y_\beta = \begin{pmatrix} \cosh\beta & 0 & \sinh\beta \\ 0 & 1 & 0 \\ \sinh\beta & 0 & \cosh\beta \end{pmatrix}, R_\alpha = \begin{pmatrix} 1 & & 0 \\ 0 & \cos\alpha & -\sin\alpha \\ 0 & \sin\alpha & \cos\alpha \end{pmatrix}.$$

From these, one obtains the matrices λ with $\det(\lambda) = 1$ and $\lambda_{00} < 0$ by multiplication with $\mathrm{diag}(-1,-1,1)$, those with $\det(\lambda) = -1$ and $\lambda_{00} > 0$ by multiplication with $\mathrm{diag}(1,-1,1)$, and finally, those with $\det(\lambda) = -1$ and $\lambda_{00} < 0$ by multiplication with $\mathrm{diag}(-1,1,1)$.

More generally $O(k,l) \subset GL(k+l,\mathbb{R})$ is the linear Lie group defined by $n = k+l$ and

$$\Gamma = \mathrm{diag}(-1,\overset{k}{\cdots},-1,+1,\overset{l}{\cdots},+1); \tag{A.6}$$

hence elements of $O(k,l)$ are matrices $\gamma \in GL(k+l,\mathbb{R})$ that satisfy $\gamma^T g \gamma = g$. Of course, the Lorentz group $O(1,3)$ is crucial for special and general relativity, but apart from $k = 1$ we will also be interested in $k = 0$ and $k = 2$. We write $O(l)$ for $O(0,l)$, which is compact, but none of the groups $O(k,l)$ with $k > 0$ is compact, except when $l = 0$, in which case $O(k,0) = O(k)$. Each group $O(l)$ has two components (distinguished as for $l = 3$ by the sign of their determinant, or, equivalently, by being orientation-preserving or reversing), whereas each $O(k,l)$ with $k > 0$ and $l > 0$ has four, distinguished by their containing I, $-I$, Γ (time reversal), and $-\Gamma$ (parity).

The additive (and hence abelian) groups $\mathbb{R}^n$ are also real linear Lie groups (although this is not their simplest description!), since one may identify $a \in \mathbb{R}^n$ with the $2n \times 2n$-matrix

$$a \equiv \begin{pmatrix} 1_n & \mathrm{diag}(a) \\ 0 & 1_n \end{pmatrix}, \tag{A.7}$$

where $\mathrm{diag}(a)$ is the diaginal $n \times n$ matrix with entries $(a_1,\ldots,a_n)$ on the diagonal. Indeed, matrix multiplication reproduces addition in $\mathrm{diag}(a)$. The last Lie groups of interest to us are

$$E(n) = O(n) \ltimes \mathbb{R}^n; \tag{A.8}$$
$$P(n) = O(1,n-1) \ltimes \mathbb{R}^n, \tag{A.9}$$

called the ***Euclidean group*** and the ***Poincaré group*** in dimension n. They are the isometry groups of the Euclidean metric $\delta = \mathrm{diag}(1,\ldots,1)$ and Minkowski metric $\eta = \mathrm{diag}(-1,1,\ldots,1)$ on $\mathbb{R}^n$, respectively. These are examples of ***semidirect products***, which are defined more generally as follows. Let some group L act linearly on a vector space V. Then the operation

$$(\lambda,v)\cdot(\lambda',v') = (\lambda\lambda', v + \lambda\cdot v'); \tag{A.10}$$
$$(\lambda,v)^{-1} = (\lambda^{-1}, -\lambda^{-1}\cdot v), \tag{A.11}$$

turns $L \ltimes V$ into a group, called the ***semidirect product*** of L and V. If $L \subset GL_n(\mathbb{R})$ is a linear Lie group and $V = \mathbb{R}^n$, then $L \ltimes V$ is a linear Lie group in $GL_{2n}(\mathbb{R})$, realized by the matrices

$$\begin{pmatrix} L & \mathbf{v} \\ 0 & 1_n \end{pmatrix}, \tag{A.12}$$

where $\mathbf{v} \in GL_n(\mathbb{R})$ is the matrix with $v \in V$ in every column.

A.2 Lie algebras

Abstractly, a ***Lie algebra*** over $\mathbb{R}$ is defined as a real vector space A equipped with an bilinear map $[\cdot,\cdot] : A \times A \to A$ that satisfies antisymmetry and the ***Jacobi identity***, i.e.,

$$[a,b] + [b,a] = 0; \tag{A.13}$$

$$[a,[b,c]] + [c,[a,b]] + [b,[c,a]] = 0 \ \ (a,b,c \in A). \tag{A.14}$$

Concretely, any linear subspace $\mathfrak{g} \subset M_n(\mathbb{R})$ that is closed under the commutator

$$[A,B] := AB - BA, \tag{A.15}$$

which automatically satisfies the Jacobi identity, is a Lie algebra (and similarly over the complex numbers). A special case is the Lie algebra of a linear Lie group $G \subset GL_n(\mathbb{R})$, subtly defined by

$$\mathfrak{g} = \{A \in M_n(\mathbb{R}) \mid e^{tA} \in G \forall t \in \mathbb{R}\}, \tag{A.16}$$

where the exponential map $\exp : \mathfrak{g} \to G$ is just given by its usual (norm-convergent) power series. It is a nontrivial fact that this concrete Lie algebra is also an abstract one, notably that $\mathfrak{g}$ is a vector space and that the bracket (A.15) indeed maps $\mathfrak{g} \times \mathfrak{g}$ to $\mathfrak{g}$. The former property follows from the ***Lie product formula***

$$e^{A+B} = \lim_{m\to\infty} \left(e^{A/m} e^{B/m}\right)^m, \tag{A.17}$$

combined with the axiom that G be closed in $GL_n(\mathbb{R})$. The latter property derives from

$$[A,B] = \frac{d}{dt} e^{tA} B e^{-tA}, \tag{A.18}$$

combined with a lemma about matrices showing that if $g \in G$ and $A \in \mathfrak{g}$, then $gAg^{-1} \in \mathfrak{g}$, which in turn follows from the definition of the exponential, implying $\exp(gAg^{-1}) = g\exp(A)g^{-1}$.

If $G = G_\Gamma$ is defined by (A.4), then its Lie algebra is

$$\mathfrak{g}_\Gamma = \{A \in M_n(\mathbb{R}) \mid A^T\Gamma = -\Gamma A\}. \tag{A.19}$$

For example, taking $\Gamma = \mathrm{diag}(1,1,1)$, the Lie algebra $\mathfrak{so}(3)$ of $SO(3)$ consists of all real 3×3 matrices X that satisfy $X^T = -X$. As a vector space $\mathfrak{so}(3) \cong \mathbb{R}^3$, since $\mathfrak{so}(3)$ has a basis

$$J_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}, \qquad J_2 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix}, \qquad J_3 = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \tag{A.20}$$

whose linear span gives all 3×3 real antisymmeric matrices. A vector space isomorphism $\mathbb{R} \xrightarrow{\cong} \mathfrak{so}(3)$ is then given by $(x,y,z) \mapsto xe_1 + ye_2 + ze_3$. The commutators of these elements are

$$[J_1,J_2] = J_3; \qquad [J_3,J_2] = -J_1; \qquad [J_3,J_1] = J_2, \tag{A.21}$$

and by linearity these determine the Lie bracket of arbitrary elements of $\mathfrak{so}(3)$.
Similarly, according to (A.19) the Lie algebra of $SO(1,2)$ consists of all $A \in M_3(\mathbb{R})$ that satisfy

$$A^T \mathrm{diag}(-1,1,1) = -\mathrm{diag}(-1,1,1)A. \tag{A.22}$$

There are good reasons for taking the basis

$$e_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}, \qquad e_2 = \begin{pmatrix} 0 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix}, \qquad e_3 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \tag{A.23}$$

with commutation relations

$$[e_1, e_2] = e_3; \qquad [e_3, e_1] = e_2; \qquad [e_3, e_2] = e_1. \tag{A.24}$$

For later use, also another basis

$$e'_1 = e_3; \qquad e'_2 = -e_2; \qquad e'_3 = e_1 \tag{A.25}$$

is useful, with Lie brackets

$$[e'_1, e'_2] = -e_3; \qquad [e'_3, e'_1] = e'_2; \qquad [e'_3, e'_2] = -e'_1. \tag{A.26}$$

Although $SO(2,1)$ is isomorphic to $SO(1,2)$, its Lie algebra has a different basis, e.g.

$$f_1 = \begin{pmatrix} 0 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix}, \qquad f_2 = \begin{pmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \qquad f_3 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \tag{A.27}$$

with commutation relations

$$[f_1, f_2] = -f_3; \qquad [f_3, f_1] = f_2; \qquad [f_3, f_2] = f_1. \tag{A.28}$$

The last interesting three-dimensional cases are the Lie algebras of the groups (A.8) and (A.9) in $n = 2$. To find the Lie brackets in a suitable basis, we note that in general the Lie algebra $\mathfrak{g}$ of a semidirect product $L \ltimes \mathbb{R}^n$ is $\mathfrak{l} \oplus \mathbb{R}^n$ as a vector space, with commutators given by

$$[(A, v), (B, w)] = ([A, B], Aw - Bv), \tag{A.29}$$

where $A, B \in \mathfrak{l}$ and $v, w \in V$. Since $SO(2)$ consists of all matrices

$$\begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix}, \quad \alpha \in [0, 2\pi], \tag{A.30}$$

we may take the basis

$$j_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \qquad j_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \qquad j_3 = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \tag{A.31}$$

the former forming a basis of $\mathbb{R}^2$, and find the commutation relations from (A.29) to be

$$[j_1, j_2] = 0; \qquad [j_3, j_1] = j_2, \qquad [j_3, j_2] = -j_1;. \tag{A.32}$$

For the Poincaré-group in $2d$, i.e. $P(2)$, on the other hand, we take

$$k_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \qquad k_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \qquad k_3 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \tag{A.33}$$

to obtain the commutation relations

$$[k_1, k_2] = 0; \qquad [k_3, k_1] = k_2; \qquad [k_3, k_2] = k_1. \tag{A.34}$$

A.3 Homogeneous manifolds

Spaces of constant curvature are special cases of homogeneous manifolds (and more specifically of symmetric spaces). To start, we quote the following basic technical result without proof:[650]

Proposition A.1 *Let G be a Lie group and $H \subset G$ a closed subgroup. Then H is itself a Lie group and there exists a smooth structure on the* **homogeneous space** *G/H such that:*

1. $\dim(G/H) = \dim(G) - \dim(H)$;
2. *The canonical projection $G \to G/H$, $\gamma \mapsto \gamma H$, is smooth;*
3. *The canonical G-action $G \times (G/H) \to G/H$, $(\gamma_1, \gamma_2 H) \mapsto (\gamma_1\gamma_2)H$, is smooth.*

We write such group actions as $\gamma_1(\gamma_2 H) = (\gamma_1\gamma_2)H$. It is clear that G acts transitively on G/H (for any $x \in G/H$ and $y \in G/H$ there is $\gamma \in G$ such that $y = \gamma x$). Without loss of generality,[651] we may also assume that G acts *effectively* on G/H: if $\gamma x = x$ for all $x \in G/H$, then $x = e$.

Homogeneous spaces arise naturally if a Lie group G acts smoothly and transitively on a manifold M (in which case M is called a ***homogeneous G-space***). Then $M \cong G/H$ with $H = G_{x'}$ (i.e. the stability group of some fixed $x' \in M$), under the diffeomorphism $M \to G/H$, $x \mapsto \gamma H$, where $\gamma \in G$ satisfies $\gamma x' = x$; the inverse map $G/H \to M$ is $\gamma H \mapsto \gamma x'$ (both maps are independent of the choice of $\gamma \in \gamma H$), and this identification $M \leftrightarrow G/H$ is G-equivariant.

The following isomorphism will be very useful in all that follows:

$$T_H(G/H) \cong \mathfrak{g}/\mathfrak{h}, \tag{A.35}$$

where $\mathfrak{g}$ and $\mathfrak{h}$ are the Lie algebras of the Lie groups G and H, respectively. To see this, let us consider a more general situation, where a Lie group G acts smoothly on a manifold M, that is, $\varphi : G \times M \to M$ is a smooth G-action on M. We will write $\varphi_\gamma(x)$ (or simply $\gamma \cdot x$) for $\varphi(\gamma, x)$, so that each map $\varphi_\gamma : M \to M$ is a diffeomorphism. For each $A \in \mathfrak{g}$ we define a map

$$A_M : C^\infty(M) \to \mathrm{Der}(C^\infty(M)); \qquad A_M f(x) = \frac{d}{dt} f(e^{tA} \cdot x)_{|t=0}. \tag{A.36}$$

This defines a derivation on $C^\infty(M)$ and hence a vector field on M, so that $A_M \in \mathfrak{X}(M)$, and we have a map $A \mapsto A_M$ from $\mathfrak{g}$ to $\mathfrak{X}(M)$. It can be shown that our map has good properties:[652]

Proposition A.2 *The map $A \mapsto A_M$ is linear and for all $A, B \in \mathfrak{g}$ satisfies*

$$[A_M, B_M] = -[A, B]_M. \tag{A.37}$$

In other words, our map is an anti-homomorphisms of Lie algebras (with respect to the usual commutator bracket of vector fields). Clearly, at any $x \in M$ we obtain a map $\mathfrak{g} \to T_x M$ by regarding $A_M(x)$ as an element of $T_x M$. In the case $M = G/H$ at hand, let us now take $x = H$.

Lemma A.3 *The linear map $\mathfrak{g} \mapsto T_H(G/H)$ defined by (A.36) has kernel $\mathfrak{h}$.*

---

[650] See e.g. Kobayashi & Nomizu (1963), Proposition I.4.2, or Helgason (1978), §II.4.

[651] If G does not act effectively on G/H, take the largest normal subgroup $H_0 \subset H$ that is also normal in G, and define $G^* = G/H_0$ and $H^* = H/H_0$. Then $G/H \cong G^*/H^*$ and G^* acts effectively on G^*/H^*. An example where this is necessary occurs if $H \subset Z(G)$, in which case all of H acts trivially on G/H. Fortunately, the isometry group of a (semi) Riemannian manifold always acts effectively on M.

[652] See e.g. Marsden & Ratiu (1994), Proposition 9.3.6. Note that $A_M = -\varphi_*(A)$, cf. (8.246).

Proof. If $A \in \mathfrak{h}$, then $\exp(tA) \in H$ by definition of $\mathfrak{h}$ (see §A.2). But $hH = H$ for any $h \in H$, whence $A_{G/H}(H) = 0$. Hence $\mathfrak{h}$ lies in the kernel of the map $\mathfrak{g} \mapsto T_H(G/H)$. Conversely, $\gamma H = H$ iff $\gamma \in H$, and $h \in H$ lies near the identity iff $h = \exp(tA)$ for some $A \in \mathfrak{h}$. □

Lemma A.3 implies (A.35) by a dimension count based on Proposition A.1.1, which gives

$$\dim(G/H) = \dim(G) - \dim(H) = \dim(\mathfrak{g}) - \dim(\mathfrak{h}) = \dim(\mathfrak{g}/\mathfrak{h}). \tag{A.38}$$

The isomorphism (A.35) gets more body of we combine it with the residual H-action on $T_H(G/H)$. For any diffeomorphism φ of M, the derivative φ'_x maps T_xM linearly to $T_{\varphi(x)}M$. If $\varphi(x) = x$, then $\varphi'_x \in \mathrm{Hom}(T_xM)$. If the diffeomorphisms φ come from a G-action on M, then

$$G_x = \{\gamma \in G \mid \gamma \cdot x = x\}. \tag{A.39}$$

is the ***stabilizer*** of x. If $\gamma \in G_x$, the linear maps $\varphi'_\gamma : T_xM \to T_xM$, combine into a homomorphism

$$\pi_x : G_x \to GL(T_xM); \qquad \gamma \mapsto \varphi'_\gamma, \tag{A.40}$$

called the ***isotropy representation*** of G_x in T_xM (here $GL(T_xM)$ consists of all invertible linear maps from T_xM to T_xM). This applies in particular to $M = G/H$ and $x = H$, so that we obtain

$$\pi_H : H \to GL(T_H(G/H)); \qquad k \mapsto \varphi'_k. \tag{A.41}$$

We will now explicitly find π_H under the isomorphism (A.35). We know that any group G acts on itself by the ***adjoint action*** $\mathrm{Ad}_\gamma(\delta) = \gamma\delta\gamma^{-1}$. If G is a Lie group,[653] this action defines a representation Ad' of G on its Lie algebra $\mathfrak{g}$, defined by $\mathrm{Ad}'_\gamma(X) = \gamma X \gamma^{-1}$. This action may, of course, be restricted to $H \subset G$, and it is easy to see that this restriction quotients to $\mathfrak{g}/\mathfrak{h}$. In our application to spaces with constant curvature, $\mathfrak{g}$ will have a ***reductive decomposition***

$$\mathfrak{g} = \mathfrak{h} \oplus \mathfrak{p}, \tag{A.42}$$

where (trivially) not only $\mathfrak{h}$, but also $\mathfrak{p}$ is invariant under Ad'_k for any $k \in H$ (if H is connected, this is equivalent to $[\mathfrak{h}, \mathfrak{p}] \subset \mathfrak{p}$). In that case, we may replace the isomorphism (A.35) by

$$T_H(G/H) \cong \mathfrak{p}, \tag{A.43}$$

Proposition A.4 *1. Under the isomorphism (A.35), the isotropy representation (A.41) of H on $T_H(G/H)$ maps to the adjoint action of H on $\mathfrak{g}/\mathfrak{h}$ (still denoted by Ad'):*

$$\pi_H(k)[A] = [\mathrm{Ad}'_k(A)], \tag{A.44}$$

where $A \in \mathfrak{g}$ and $[A] \in \mathfrak{g}/\mathfrak{h}$, seen as an element of $T_H(G/H)$ via the isomorphism (A.35).

2. Consequently, under the isomorphism (A.43), assuming that $\mathfrak{p}$ is $\mathrm{Ad}'(H)$-invariant, the same isotropy representation of H is mapped to the adjoint action of H on $\mathfrak{p}$.

Indeed, for any $A \in \mathfrak{g}$, $k \in H$, and $f \in C^\infty(G/H)$ we have, cf. (A.36) and (A.41),

$$\begin{aligned}(\pi_H(k)A_{G/H})f(H) &= \frac{d}{dt} f(ke^{tA} \cdot H)_{|t=0} = \frac{d}{dt} f(ke^{tA}k^{-1} \cdot H)_{|t=0} \\ &= \frac{d}{dt} f(e^{tkAk^{-1}} \cdot H)_{|t=0} = (\mathrm{Ad}'_k A)_{G/H} f(H).\end{aligned}$$ □

[653] It follows from our definition of a Lie algebra in Appendix A.2 that Ad' is well defined as well as linear.

The following examples of homogeneous spaces arose in §4.4, restricted to dimension two:

$$S^2 \cong O(3)/O(2); \qquad dS^2 \cong O(1,2)/O(1,1); \tag{A.45}$$
$$\mathbb{R}^2 \cong E(2)/O(2); \qquad \mathbb{R}^2 \cong P(2)/O(1,1); \tag{A.46}$$
$$H^2 \cong O(1,2)/O(2); \qquad AdS^2 \cong O(2,1)/O(1,1), \tag{A.47}$$

where we put $\rho = 1$ (and also $S^2_1 \equiv S^2$, etc.), so that the non-flat spaces in question are given by

$$S^2 = \{(x_1,x_2,x_3) \in \mathbb{R}^3 \mid x_1^2 + x_2^2 + x_3^2 = 1\}; \tag{A.48}$$
$$dS^2 = \{(x_0,x_1,x_2) \in \mathbb{R}^3 \mid -x_0^2 + x_1^2 + x_2^2 = 1\}; \tag{A.49}$$
$$H^2 = \{(x_0,x_1,x_2) \in \mathbb{R}^3 \mid -x_0^2 + x_1^2 + x_2^2 = -1\}; \tag{A.50}$$
$$AdS^2 = \{(x_{-1},x_0,x_1) \in \mathbb{R}^3 \mid -x_{-1}^2 - x_0^2 + x_1^2 = -1\}, \tag{A.51}$$

and the Lie groups in question were defined in Appendix A.1. In $d = 2$ de Sitter space dS^2 is diffeomorphic to anti de Sitter space AdS^2, but they will be different as Lorentzian manifolds, as in the case $\mathbb{R}^2$ in (A.46). To verify (A.45), let $O(3)$ act on S^2 by restricting its defining action on $\mathbb{R}^3$, and take $x' \in S^2$ to be the north pole $(0,0,1)$, in which case the $O(2)$ in (A.45) consists of rotations around the z-axis and reflections in planes through the origin that contain the z-axis.[654] Similarly, for dS^2, where one also takes $x' = (0,0,1)$, and for H^2 and AdS^2, where the most convenient fiducial point is $(1,0,0)$. For (A.46), we let the $2d$ Euclidean group $E(2)$ and the $2d$ Poincaré group $P(2)$ act on $\mathbb{R}^2$ in the defining representation, and take $x' = (0,0)$.

Writing (A.45) - (A.47) generically as $M \cong G/H$, where $H = O(2)$ or $H = O(1,1)$, the $\mathrm{Ad}'(H)$-invariant decomposition (A.42) applies to each G in the list. In all six cases we have

$$\mathfrak{g} \cong \mathbb{R}^3, \qquad \mathfrak{h} \cong \mathbb{R}, \qquad \mathfrak{p} \cong \mathbb{R}^2, \tag{A.52}$$

as vector spaces, taking $\mathfrak{h}$ to be the linear span of the third generator and $\mathfrak{p}$ to be the linear span of the first two generators: see (A.21) for the Lie algebra of $O(3)$ as relevant for S^2, see (A.24) for $SO(1,2)$ in the context of dS^2, see (A.26) again for $SO(1,2)$ but now applied to H^2, then (A.27) for $O(2,1)$ applied to AdS^2, then (A.31) for $E(2)$ applied to $\mathbb{R}^2$ in Riemannian signature, and finally, eq. (A.34) for $P(2)$ applied to $\mathbb{R}^2$ in Lorentzian signature. All cases give

$$[\mathfrak{h},\mathfrak{p}] \subset \mathfrak{p}. \tag{A.53}$$

Lemma A.5 *In all six cases the decomposition* (A.42) *is* $\mathrm{Ad}'(H)$*-invariant. In addition, under the last isomorphism in* (A.52) *the adjoint* H*-action on* $\mathfrak{p}$ *is just its defining action on* $\mathbb{R}^2$.

The significance of this observation will become clear in the next section. The proof is long!

Proof. Let $u : G \to GL(V)$ be a representation (i.e. a homomorphism) of a Lie group G on a *finite-dimensional* vector space V. For $A \in \mathfrak{g}$ we define a linear map $du(A) : V \to V$ by

$$du(A)v = \frac{d}{dt} u\left(e^{tA}\right) v_{|t=0}. \tag{A.54}$$

[654] Note that $O(2)$ has two components, like $O(3)$, again distinguished by $\det = \pm 1$. Elements γ with $\det(\gamma) = 1$ are rotations whereas those with $\det(\gamma) = -1$ are reflections in a line through the origin (e.g. $\mathrm{diag}(1,-1)$).

Letting $A \in \mathfrak{g}$ vary, this construction gives a linear map $du : \mathfrak{g} \to \mathrm{Hom}(V)$, which satisfies

$$[du(A), du(B)] = du([A,B]); \qquad e^{du(A)} = u\left(e^A\right). \tag{A.55}$$

In particular, if G is connected, then u can be recovered from du via (A.55). If G is simply connected, this even gives an equivalence between finite-dimensional Lie group and Lie algebra representations. For example, the adjoint representation $\mathrm{Ad}' : G \to GL(\mathfrak{g})$, $\mathrm{Ad}'(\gamma)A = \gamma A \gamma^{-1}$, defines a Lie algebra homomorphism $\mathrm{ad} : \mathfrak{g} \to \mathrm{Hom}(\mathfrak{g})$,[655] where $\mathrm{ad} \equiv d(Ad')$, namely

$$\mathrm{ad}(A)B = [A,B]. \tag{A.56}$$

In view of this, for $G = O(3)$, the commutation relations (A.21) show that

$$\mathrm{ad}(J_3)J_1 = J_2; \qquad \mathrm{ad}(J_3)J_2 = -J_1, \tag{A.57}$$

where we repeat that J_3 is the generator of the subgroup $O(2)$ of $O(3)$ that consists of rotations around the z-axis. This means that as a matrix relative to the basis (J_1, J_2) of $\mathbb{R}^2$, the restriction of the linear map $\mathrm{ad}(J_3) : \mathfrak{so}(3) \to \mathfrak{so}(3)$ to $\mathfrak{p} = \mathrm{span}(J_1, J_2) \cong \mathbb{R}^2$ (which restriction is well defined, as the above relations show) is just the usual generator of $\mathfrak{so}(2)$, see (A.31), which is obtained from the defining action id of $G = O(2)$ on $V = \mathbb{R}^2$ by the procedure (A.54). By exponentiation, we then conclude that the corresponding Ad-action of $SO(2)$ on $\mathfrak{p}$ is the defining action, too. To obtain the Ad-action of all of $O(2)$ it suffices to take the element

$$R_x = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \text{ which, seen as an element of } O(3), \text{ is } R_{xz} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

the reflection in the x-z plane. Its adjoint action on the generators of $\mathbb{R}^2$ can be computed to give

$$\mathrm{Ad}_{R_{xz}} J_1 = R_{xz} J_1 R_{xz}^{-1} = -J_1; \qquad \mathrm{Ad}_{R_{xz}} J_2 = R_{xz} J_2 R_{xz}^{-1} = J_2, \tag{A.58}$$

which means that the adjoint representation of R_x on $\mathfrak{p} = \mathrm{span}(J_1, J_2)$ is not only well defined in mapping $\mathfrak{p}$ to itself, but also that under the identification $\mathfrak{p} \cong \mathbb{R}^2$, $\mathrm{Ad}'_{|\mathfrak{p}}$ maps R_x to itself.

However, the (J_1, J_2) basis of $\mathbb{R}^2$ is not the geometrically natural basis in the given context. Instead, we compute the map (A.36), first at arbitrary points $(x_1, x_2, x_3) \in S^2$. This gives

$$J_1 \mapsto -x_3 \frac{\partial}{\partial x_2} + x_2 \frac{\partial}{\partial x_3}; \qquad J_2 \mapsto x_3 \frac{\partial}{\partial x_1} - x_1 \frac{\partial}{\partial x_3}; \qquad J_3 \mapsto -x_2 \frac{\partial}{\partial x_1} + x_1 \frac{\partial}{\partial x_2}. \tag{A.59}$$

At the point $x' = (0,0,1) \in S^2$, the generator J_3 is mapped to zero and for the other two we have

$$J_1 \mapsto -\frac{\partial}{\partial x_2}; \qquad J_2 \mapsto \frac{\partial}{\partial x_1}. \tag{A.60}$$

Hence the natural basis for $T_{x'} S^2 \cong \mathbb{R}^2$ is $u_1 = (1,0) = J_2 = \partial/\partial x_1$ and $u_2 = (0,1) = -J_1 = \partial/\partial x_2$. Fortunately, this leads to exactly the same conclusions as the previous basis, as the reader can easily verify. This proves Lemma A.5 for S^2, i.e., $G = SO(3)$ and $H = SO(2)$.

[655] Each map $\mathrm{ad}(A)$ is even a derivation of $\mathfrak{g}$ as a Lie algebra, as follows from the Jacobi identity.

For hyperbolic space H^2, that is, $G = SO(1,2)$ and $H = SO(2)$, we use the basis (e'_1, e'_2, e'_3) of the Lie algebra of $O(1,2)$ defined in (A.26) and the fiducial point $x' = (1,0,0) \in H^2$. Then

$$e'_1 \mapsto \frac{\partial}{\partial x_1}; \qquad e'_2 \mapsto \frac{\partial}{\partial x_2}, \tag{A.61}$$

in coordinates (x_0, x_1, x_2) on $\mathbb{R}^3$, so that for H^2 the right basis of $\mathbb{R}^2$, seen as $\mathfrak{p} = \mathrm{span}(e'_1, e'_2)$, is simply $u_1 = e'_1$ and $u_2 = e'_2$. The Lie brackets (A.26), now reinterpreted as

$$\mathrm{ad}(e'_3)e'_1 = e'_2; \qquad \mathrm{ad}(e'_3)e'_2 = -e'_1, \tag{A.62}$$

then lead to the same conclusion as for S^2: the Ad-action of $SO(2)$ on $\mathbb{R}^2$ is the defining action. This is also true for the component that is not connected to the identity; the matrix R_x shown above is now embedded into $SO(1,2)$ as $\mathrm{diag}(1,1,-1)$, but the result remains $\mathrm{Ad}(R_x) = R_x$.

For Euclidean $\mathbb{R}^2$, i.e. $G = E(2)$ and $H = O(2)$, $\mathfrak{p}$ is literally $\mathbb{R}^2$, seen as the Lie algebra of the second factor in the semidirect product (A.8), and the lemma should be evident from (A.32).

For de Sitter space dS^2 we need similar computations as for S^2 and H^2. Using the basis (A.24) and the fiducial point $x' = (0,0,1)$ on dS^2, we find

$$e_1 \mapsto -\frac{\partial}{\partial x_1}; \qquad e_2 \mapsto -\frac{\partial}{\partial x_0}, \tag{A.63}$$

in coordinates (x_0, x_1, x_2) on $\mathbb{R}^3$, making $u_0 = (1,0) = -e_2$ and $u_1 = (0,1) = -e_1$ the natural basis of $\mathfrak{p} \cong \mathbb{R}^2$. This gives $\mathrm{ad}(e_3)u_0 = -[e_3, e_2] = -e_1 = u_1$ and $\mathrm{ad}(e_3)u_1 = -[e_3, e_1] = -e_2 = u_0$. which implies that $\mathrm{ad}(e_3)$ is the matrix k_3 in (A.33), coming from the $2d$ boost generator

$$\begin{pmatrix} \cosh\chi & \sinh\chi \\ \sinh\chi & \cosh\chi \end{pmatrix}, \ \chi \in \mathbb{R}. \tag{A.64}$$

Thus the adjoint k_3-action on $\mathfrak{p}$ generates the defining $O(1,1)_0$-action on $\mathbb{R}^2$. This time there are three other components that contribute to the full $O(1,1)$-action, generated by the matrices

$$P = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \qquad T = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad PT = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}, \tag{A.65}$$

which as elements of $O(1,2)$ under $O(1,1) \subset O(1,2)$ are given by

$$\tilde{P} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad \tilde{T} = \begin{pmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad \widetilde{PT} = \begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \tag{A.66}$$

Then $\mathrm{Ad}(P)u_0 = -\tilde{P}e_2\tilde{P}^{-1} = -e_2 = u_0$ and $\mathrm{Ad}(P)u_1 = -\tilde{P}e_1\tilde{P}^{-1} = e_1 = -u_1$, which show that $\mathrm{Ad}(P) = P$ on $\mathfrak{p} \cong \mathbb{R}^2$. The other two cases are similar, which settles the case of dS^2.

For AdS^2, in coordinates (x_{-1}, x_0, x_1) with fiducial point $x' = (1,0,0)$, we obtain

$$f_1 \mapsto -\frac{\partial}{\partial x_1}; \qquad f_2 \mapsto -\frac{\partial}{\partial x_0}, \tag{A.67}$$

which is similar to (A.63). Indeed, using the basis $u_0 = (1,0) = -f_2$ and $u_1 = (0,1) = -f_1$, and the Lie brackets (A.28), we obtain $\mathrm{ad}(f_3)u_0 = -[f_3, f_2] = -f_1 = u_1$ and likewise $\mathrm{ad}(f_3)u_1 = -[f_3, f_1] = -f_2 = u_0$, so that once again $\mathrm{ad}(f_3)$ is given by the matrix k_3 in (A.33), which generates the $2d$ boosts in (A.64). We leave the verification that the discrete elements (A.65) of $O(1,1)$ also act correctly on $\mathfrak{p} = \mathrm{span}(f_1, f_2)$ to the reader; their embedding in $O(2,1)$ is different from (A.66), and is now given by always having $+1$ in the upper left entry.

Finally, the case of Minkowski $\mathbb{R}^2$, with $G = P(2)$ acting on $\mathbb{R}^2$ and $H = O(1,1)$, is very similar to the Euclidean case, with eqs. (A.33) - (A.34) replacing (A.31) - (A.32). □

A.4 Symmetric spaces

So far, this was an exercise in Lie group theory and differential geometry. We now bring in a metric. The relationship between homogeneous spaces and metric geometry is twofold:

1. Given $M = G/H$ (always connected), one may study possible G-invariant metrics g on M.
2. Given (M,g), one may find out if the isometry group of g possibly acts transitively on M.

In general, a metric g on M is *invariant* under a diffeomorphism φ of M if

$$\varphi^* g = g \qquad \Leftrightarrow \qquad g_{\varphi(x)}(\varphi'_x(X), \varphi'_x(Y)) = g_x(X,Y) \ \ \forall x \in M, X, Y \in T_x M. \tag{A.68}$$

The set of all such diffeomorphisms φ is the ***isometry group*** of (M,g), denoted by $\mathrm{Iso}(M,g)$. If M is a G-space, we say that g is G-invariant if $\varphi_\gamma^* g = g$ for all $\gamma \in G$. If this is the case, then $G \subset \mathrm{Iso}(M,g)$ by definition (typically without equality). If in addition G acts transitively on M, we say that (M,g) is a ***homogeneous (semi) Riemannian manifold***, so that $M \cong G/H$.

We return to the second point in the next section. The first is settled as follows:

Proposition A.6 *1. There is a bijective correspondence between G-invariant semi-Riemannian metrics on G/H and $\mathrm{Ad}'(H)$-invariant metrics on $\mathfrak{g}/\mathfrak{h}$ (in the sense of Definition 2.5) and hence, if (A.42) applies, on $\mathfrak{p}$.*

2. If $H \cong G_\Gamma$ for some $G_\Gamma \subset GL_n(\mathbb{R})$ as defined in (A.4), and the $\mathrm{Ad}'(H)$-action on $\mathfrak{g}/\mathfrak{h}$ (or, if applicable, on $\mathfrak{p}$) is equivalent to the defining action of G_Γ on $\mathbb{R}^n$, then $\mathfrak{g}/\mathfrak{h}$ (etc.) has a unique $\mathrm{Ad}'(H)$-invariant metric (up to scaling by a nonzero constant), and hence G/H has a a unique G-invariant semi-Riemannian metric (up to scaling by a nonzero constant).

Proof. To prove the first claim, just use (A.35) or (A.43): any metric on $\mathfrak{g}/\mathfrak{h}$ or $\mathfrak{p}$ defines a metric g on $T_H(G/H)$, which the G-action then pushes to any other point. Invariance under G clearly requires $\varphi_k^* g_H = g_H$ for any $k \in H$, so that Proposition A.4 shows that $\mathrm{Ad}'(H)$-invariance of the inner product is necessary. It is a simple exercise to show that it is also sufficient.[656]

For the second claim, any metric g on $\mathbb{R}^n$ takes the form

$$g(x,y) = \langle x, Ay \rangle_\Gamma = \langle x, \Gamma A y \rangle, \tag{A.69}$$

for some $A \in GL_n(\mathbb{R})$ (to see this, regard metrics as symmetric quadratic forms). G_Γ-invariance of $\langle -,- \rangle_\Gamma$ gives $\langle \gamma x, y \rangle_\Gamma = \langle x, \gamma^{-1} y \rangle_\Gamma$, so that G_Γ-invariance of g, i.e. $g(\gamma x, \gamma y) = g(x,y)$ for all $x, y \in \mathbb{R}^n$ and $\gamma \in G_\Gamma$, is equivalent to $[A, \gamma] = 0$ for all $\gamma \in G_\Gamma$. Since the G_Γ-action on $\mathbb{R}^n$ is irreducible, Schur's lemma gives $A = \lambda \cdot \mathrm{id}$, for some $\lambda \neq 0$, so that $g(x,y) = \lambda \langle x, y \rangle_\Gamma$. □

Proposition A.7 *For any Riemannian or Lorentzian manifold (M,g) and $G \subset \mathrm{Iso}(M,g)$, the isotropy representation $\pi_x(G_x)$ defined in (A.40) is injective.*

Proof. Near x, any isometry φ of M is determined by its tangent map φ'_x at some fixed $x \in M$: to find $\varphi(y)$ for y in a normal nbhd U_x, assume $y = \exp_x(Y)$ for some $Y \in T_x M$. If φ is an isometry, then $\varphi(\exp_x(Y)) = \exp_x(\varphi'_x(Y))$. Injectivity of π_x then follows from (A.40). □

[656] See e.g. Proposition 3.1 in Kobayashi & Nomizu (1969) or Proposition 11.22 in O'Neill (1983).

Our proof of Corollary 4.11 in the main text is based on the concept of a ***symmetric space***.[657]

Definition A.8 *1. A (semi) Riemannian manifold (M,g) is* **locally symmetric** *if each $x \in M$ has a normal nbhd U_x and an isometry $l_x : U_x \to U_x$ with the following properties:*

$$l_x(x) = x \qquad\qquad (l_x)'_x = -\mathrm{id}_{T_xM}. \tag{A.70}$$

2. It is called **symmetric** *if, for each $x \in M$, the above properties hold for $U_x = M$,*

Such a map l_x is often called a ***geodesic reflection***, since (A.70) is equivalent to

$$l_x(\exp_x(X)) = \exp_x(-X), \qquad\qquad X \in \exp_x^{-1}(U_x) \subset T_xM. \tag{A.71}$$

Eq. (A.71), and hence also (A.70), gives $l_x^2 = \mathrm{id}_{U_x}$. Eq. (A.71) easily implies (A.70), and the converse implication follows from the fact that, as just mentioned, near x a local isometry φ is determined by its tangent φ'_x at x. In view of the assumptions in Corollary 4.11, we note:[658]

Lemma A.9 *If (M,g) is complete and simply connected and is locally symmetric, then it is symmetric. Conversely, a symmetric space is complete.*

The connection between symmetric spaces and spaces with constant curvature will run via:

Lemma A.10 *A space (M,g) is locally symmetric iff $\nabla\mathrm{Riem} = 0$.*

The implication "$\Rightarrow$" is a simple exercise. For the converse, take $x, y \in M$ and let $F : T_xM \to T_yM$ be a linear isomorphism. If U_x and U_y are normal nbhds of x and y, we obtain a map

$$f : U_x \to U_y; \qquad\qquad f := \exp_y \circ F \circ \exp_x^{-1}. \tag{A.72}$$

It follows from the ***Cartan–Ambrose–Hicks theorem*** that if F preserves both the metric and the Riemann tensor, and in addition $\nabla\mathrm{Riem} = 0$, then f is an isometry.[659] If $x = y$, then $F := -\mathrm{id}_{T_xM}$ trivially satisfies the assumptions of this theorem, simply because both g and Riem have even rank (namely 2 and 4, respectively). The ensuing map f is our desired local isometry l_x. □

Proposition A.11 *The isometry group $\mathrm{Iso}(M,g)$ of a symmetric space acts transitively on M. Moreover, already its identity component $\mathrm{Iso}(M,g)_0$ acts transitively on M.*

Proof. First assume that any two points y, z of M may be connected by a geodesic γ (in the Riemannian case this is true by Lemma A.9 and the Hopf–Rinow theorem). So let $y = \gamma(0)$ and $z = \gamma(T)$. Then $y = l_x(z)$ for $x = \gamma(T/2)$, and we recall that l_x is an isometry. In general, the same argument applies to each segment of a chain of geodesic segments connecting y and z. This argument can be iterated to connect y to z via a composition of arbitrarily many small geodesic reflections, each contained in $\mathrm{Iso}(M,g)_0$, which yields the second claim. □

[657] See Helgason (1978), *passim*, Kobayashi & Nomizu (1969), chapter IX, and Joos (2002), chapter 5.

[658] See Kobayaski & Nomizu (1969), Corollary VI.7.9 and Theorems XI.1.2 and 1.3.

[659] See Kobayaski & Nomizu (1963), Theorem 7.4. The Cartan–Ambrose–Hicks theorem states that f is a (local) isometry iff F preserves g and Riem, and for all $Y \in T_xM$ such that $\exp_x(Y) \in U_x$ one has

$$\mathrm{Riem}_{\exp_x(F(Y))}(P_Y(U), P_Y(V), P_Y(W), P_Y(X)) = \mathrm{Riem}_{\exp_x(Y)}(U,V,W,X)$$

for all $U,V,W,X \in T_{\exp_x(Y)}$, where $P_Y : T_{\exp_x(Y)}M \to T_xM \to T_{\exp_x(F(Y))}M$ is the composition of parallel transport along the geodesics γ_Y (traversed backward) and $\gamma_{F(Y)}$. This condition is automatically satisfied when $\nabla\mathrm{Riem} = 0$.

A.5 Classification of spaces with constant curvature

Our proof of Corollary 4.11 consists of three steps, of which we state the first two as a lemma:

Lemma A.12 *1. If (M,g) has constant curvature, it is locally symmetric. Consequently, if (M,g) is simply connected, complete, and has constant curvature, then it is symmetric.*

2. If (M,g) is symmetric, then it is a homogeneous (semi) Riemannian manifold.

Therefore, among (semi) Riemannian manifolds we have the following implications:

$$\text{constant curvature} \Rightarrow \text{symmetric} \Rightarrow \text{homogeneous}.$$

This lemma reduces the classification problem of spaces with constant curvature to a problem in Lie groups and Lie algebras, which we will discuss and solve. The second part of the above lemma is a restatement of Proposition A.11. For the first part we return to the proof of Proposition 4.7, namely $\mathrm{Riem}_x = k(x)S$. Taking the covariant derivative with respect to an arbitrary vector-field $U \in \mathfrak{X}(M)$ gives $\nabla_U \mathrm{Riem} = (Uk)\cdot S$, since $\nabla_U S = 0$ by definition of the Levi-Civita connection (which gives $\nabla_U g = 0$). Eq. (4.23) then gives, for arbitrary $X,Y,Z \in \mathfrak{X}(M)$,

$$\begin{aligned}(Uk)\cdot(g(Z,Y)X - g(Z,X)Y) + (Xk)\cdot(g(Z,U)Y - g(Z,Y)U)\\ + (Yk)\cdot(g(Z,X)U - g(Z,U)X) = 0.\end{aligned} \tag{A.73}$$

The first part of the lemma then follows from Lemma A.10.[660] □

Hence under the assumptions of Corollary 4.11 we have $M \cong G/H$, with $G = \mathrm{Iso}(M,g)$ (or $G = \mathrm{Iso}(M,g)_0$), and $H = G_{x'}$ for some $x' \in M$ (or its identity component H_0). By Proposition 4.10, the given G-invariant (constant curvature) metric g on M is entirely determined by some suitable inner product $\langle\cdot,\cdot\rangle$ on $\mathfrak{g}/\mathfrak{h}$, and by Proposition A.4 the H-action on $T_{x'}M$ is mapped to the $\mathrm{Ad}'(H)$-action on $\mathfrak{g}/\mathfrak{h}$ (which by implication preserves $\langle\cdot,\cdot\rangle$). By Proposition A.7 the representation Ad' is injective on H so if we choose an orthonormal basis of $\mathfrak{g}/\mathfrak{h}$ with respect to $\langle\cdot,\cdot\rangle$, and hence obtain an identification $\mathfrak{g}/\mathfrak{h} \cong \mathbb{R}^n$, we may also identify $H \cong \mathrm{Ad}'(H)$ with a certain subgroup of $O(n)$ in the Riemannian case, or of $O(1,n-1)$ in the Lorentzian case.

Lemma A.13 *If, in the situation just described, (M,g) has constant curvature and $G = \mathrm{Iso}(M,g)$, then $H = O(n)$ in the Riemannian case and $H = O(1,n-1)$ in the Lorentzian case.*

This follows by the argument in the proof of Lemma A.10, which applies because constant curvature implies $\mathrm{Riem} = k\cdot S$, see Proposition 4.7 and especially eq. (4.86). Any element $F \in O(n)$ or $F \in O(1,n-1)$ preserves the inner product, and hence the metric, and hence, by the above formula, the Riemann tensor. Thus F comes from an isometry f, i.e. $F \in H$. □

We now know that $M \cong G/H$ as a homogeneous Riemannian or Lorentzian manifold, where

$$G = \mathrm{Iso}(M,g); \tag{A.74}$$

$$H = O(n) \text{ or } O(1,n-1). \tag{A.75}$$

[660] This argument also leads to a proof of the claim below Definition 4.6 to the effect that if M is connected, $\dim(M) \geq 3$, and $C_x(X,Y)$ is independent of X and Y for each x, then this common value is also independent of x. Indeed, in $d \geq 3$ we may take $Z = U$ to be unit vectors and (X,Y,Z) mutually perpendicular, so that (A.73) yields $(Xk)\cdot Y - (Yk)\cdot X = 0$. Since this is true for all $X \perp Y$, it follows that $Xk = Yk = 0$, and hence k is constant.

Since $O(n)$ and $O(1,n-1)$ act irreducibly on $\mathbb{R}^n$, so that $\mathrm{Ad}'(H)$ is irreducible on $\mathfrak{g}/\mathfrak{h}$, by Proposition A.6 there is exactly one possible G-invariant metric g on G/H (up to scaling).

We now transfer the involutions l_x on M to G. Since for all $x \in M$ and $\gamma \in \mathrm{Iso}(M,g)$ one has

$$\gamma l_x \gamma^{-1} = l_{\gamma x}, \tag{A.76}$$

it is sufficient to consider a single $l_{x'} : M \to M$, where $x' \in M$ is arbitrary. For (A.74), define

$$l : G \to G; \tag{A.77}$$
$$\gamma \mapsto l_{x'} \gamma l_{x'}. \tag{A.78}$$

Using (A.76) and the definition of the maps l_x, it is easy to show that l has the properties

$$l \neq \mathrm{id}_G; \qquad l^2 = \mathrm{id}_G; \qquad l(\gamma\delta) = l(\gamma)l(\delta). \tag{A.79}$$

We defined l by (A.78) for (A.74) - (A.75), in which context (A.79) follow from the definition. Conversely, for any Lie group G one may start with a nontrivial smooth involutive automorphism (A.77), i.e. a map (A.77) satisfying (A.79), called a ***Cartan involution*** on G, and define

$$H := G^l \equiv \{\gamma \in G \mid l(\gamma) = \gamma\} \tag{A.80}$$

as the fixed-point set of l. Then construct a family $(l_x)_{x \in G/H}$ of diffeomorphisms of G/H by

$$l_H(\gamma H) := l(\gamma) H; \tag{A.81}$$
$$l_{\gamma H}(x) := \gamma \cdot l_H(\gamma^{-1} \cdot x). \tag{A.82}$$

If H is connected, these procedures are equivalent; if H is disconnected, then $G^l_0 \subset H \subset G^l$.[661] Thus one may start either with a symmetric space (M,g) or with the corresponding (Lie) group-theoretical data (G,l). Up to issues with connectedness, which have to be dealt with by hand, these group-theoretical data can in turn be replaced by algebraic data, to which we now turn.

Since $l : G \to G$ is smooth, it has a derivative $l' : \mathfrak{g} \to \mathfrak{g}$, defined by, cf. (A.16),

$$l'(A) = \frac{d}{dt} l\left(e^{tA}\right)_{t=0}. \tag{A.83}$$

As in (A.55), this map satisfies $\exp(l'(A)) = l(\exp(A))$. From this, and $l^2 = \mathrm{id}_G$, we compute

$$l' \circ l'(A) = \frac{d}{dt} l\left(e^{tl'(A)}\right)_{t=0} = \frac{d}{dt} l\left(l(\exp(tA))\right)_{t=0} = \frac{d}{dt}\left(e^{tA}\right)_{t=0} = A, \tag{A.84}$$

so that $(l')^2 = \mathrm{id}_{\mathfrak{g}}$. We therefore have our promised canonical decomposition (A.42), in which $\mathfrak{h}$ and $\mathfrak{p}$ are the eigenspaces of l' with eigenvalue 1 and -1, respectively. Furthermore, it follows from the last entry in (A.79) that l' is a Lie algebra automorphism, i.e., l' is linear and

$$l'([A,B]) = [l'(A), l'(B)]. \tag{A.85}$$

This implies the following properties (of which the first one is trivial since $H \subset G$ is a subgroup):

$$[\mathfrak{h},\mathfrak{h}] \subset \mathfrak{h}; \qquad [\mathfrak{h},\mathfrak{p}] \subset \mathfrak{p}; \qquad [\mathfrak{p},\mathfrak{p}] \subset \mathfrak{h}. \tag{A.86}$$

[661] See Helgason (1978) or even wikipedia, `symmetric space`, which entry is excellent.

We return to the proof of Corollary 4.11 (our classification problem). Proposition A.6 and A.74 - (A.75) apply, as well as the remarks preceding Lemma A.13. Consequently,

$$\mathfrak{p} \cong \mathbb{R}^n, \tag{A.87}$$

and the $\mathrm{Ad}'(H)$-action on $\mathfrak{p}$ is the defining action of $H = O(n)$ or $H = O(1,n-1)$ on $\mathbb{R}^n$. By (A.56), the derivative of the $\mathrm{Ad}'(H)$-action is the $\mathrm{ad}(\mathfrak{h})$-action, i.e. for $A \in \mathfrak{h}$ and $v \in \mathfrak{p}$ we have

$$[A, v] = A \cdot v, \tag{A.88}$$

where $A \cdot v$ is the derivative of the defining action of H_0, see (A.54). Since the Lie bracket $[A,B]$ for $A, B \in \mathfrak{h}$ is also known (because $\mathfrak{h} = \mathfrak{o}(\mathfrak{n})$ or $\mathfrak{h} = \mathfrak{o}(\mathfrak{1}, \mathfrak{n}-\mathfrak{1})$), by (A.86) all we need to find out to determine $\mathfrak{g}$ as a Lie algebra (and hence, by ***Lie's third theorem***,[662] to determine G as a Lie group) is the commutator $[u,v] \in \mathfrak{h}$ of $u, v \in \mathfrak{p} \cong \mathbb{R}^n$. For $n = 2$ the only known unknown is

$$[T_1, T_2] = \rho T_3, \tag{A.89}$$

for some constant $\rho \in \mathbb{R}$, where (T_1, T_2) is some basis of $\mathbb{R}^2$ and

$$T_3 = j_3 \qquad H = O(n); \tag{A.90}$$
$$T_3 = k_3 \qquad H = O(1,1), \tag{A.91}$$

see (A.31) and (A.33), respectively. Rescaling the metric by a positive constant then restricts us to $\rho = 1, 0, -1$. For $H = O(2)$ this leaves us with the following list:

$$\rho = 1: \quad [T_1,T_2] = T_3; \quad [T_3,T_1] = T_2; \quad [T_3,T_2] = -T_1; \tag{A.92}$$
$$\rho = 0: \quad [T_1,T_2] = 0; \quad [T_3,T_1] = T_2; \quad [T_3,T_2] = -T_1; \tag{A.93}$$
$$\rho = -1: \quad [T_1,T_2] = -T_3; \quad [T_3,T_1] = T_2; \quad [T_3,T_2] = -T_1. \tag{A.94}$$

These are the Lie algebras of $O(3)$, $E(2)$, and $O(1,2)$, respectively, cf. (A.21), (A.32), and (A.26). Thus we find the homogeneous spaces

$$O(3)/O(2) \cong S^2; \qquad \rho = 1, \tag{A.95}$$
$$E(2)/O(2) \cong \mathbb{R}^2; \qquad \rho = 0, \tag{A.96}$$
$$O(1,2)/O(2) \cong H^2; \qquad \rho = -1, \tag{A.97}$$

see the left-hand sides of (A.45) - (A.47) in §A.3.

For $H = O(1,1)$ each third bracket changes sign (k_3 versus j_3), and hence we obtain

$$\rho = 1: \quad [T_1,T_2] = T_3; \quad [T_3,T_1] = T_2; \quad [T_3,T_2] = T_1; \tag{A.98}$$
$$\rho = 0: \quad [T_1,T_2] = 0; \quad [T_3,T_1] = T_2; \quad [T_3,T_2] = T_1; \tag{A.99}$$
$$\rho = -1: \quad [T_1,T_2] = -T_3; \quad [T_3,T_1] = T_2; \quad [T_3,T_2] = T_1. \tag{A.100}$$

[662] Let $\mathfrak{g}$ be a Lie algebra. There exists a simply connected Lie group $\tilde{G}$, unique up to isomorphism, such that the Lie algebra of $\tilde{G}$ is $\mathfrak{g}$ (and any Lie group isomorphic to $\tilde{G}$ has a Lie algebra ismomorphic to $\mathfrak{g}$). Furthermore, if G is a connected Lie group with Lie algebra isomorphic to $\mathfrak{g}$, then $G \cong \tilde{G}/D$, where D is a discrete normal subgroup of the center of $\tilde{G}$. This called ***Lie's third theorem*** (first proved by Cartan). See e.g. Duistermaat & Kolk (2000).

Now we have the Lie algebras of $O(1,2)$ (bis), $P(2)$, and $O(2,1)$, see (A.23), (A.34), and (A.28), respectively, and therewith, the homogeneous spaces

$$O(1,2)/O(1,1) \cong dS^2; \qquad \rho = 1, \tag{A.101}$$

$$P(2)/O(1,1) \cong \mathbb{R}^2; \qquad \rho = 0, \tag{A.102}$$

$$O(2,1)/O(1,1) \cong AdS^2; \qquad \rho = -1, \tag{A.103}$$

see the right-hand sides of (A.45) - (A.47) in §A.3.

We still need to compute the (constant) curvature of these spaces.

Lemma A.14 *For any symmetric space $M = G/H$, the Riemann tensor at $H \in G/H$ is given by*

$$\mathrm{Riem}_H(X,Y,X,Y) = -g_H([[X,Y],Y],X), \tag{A.104}$$

where g_H is the metric at $H \in G/H$ (which determines the metric on G/H, cf. Proposition A.6).

Proof. This formula follows from (4.12) and the Koszul formula (3.54) for the covariant derivative of the Levi-Civita connection.[663] It is enough to verify (A.104) for orthonormal basis vectors $X = T_a$ and $Y = T_b$, which come from a basis of $\mathfrak{p}$, as explained in the main text for $n = 2$. In (3.54) only the last three (commutator) terms are nonzero, whilst in (4.12) only the term $-g(X, \nabla_{[X,Y]})Y)$ contributes; the others all involve commutators taking values in $\mathfrak{h}$, which give vectors that vanish at $H \in G/H$. This gives

$$R_{abab} = -\tfrac{1}{2}(g_H([[T_a,T_b],T_b],T_a) + g([[T_a,[T_a,T_b]],T_b)). \tag{A.105}$$

These two terms are equal because of $\mathrm{ad}(\mathfrak{h})$-invariance of g_H, which gives

$$g_H([X,Y],Z) + g(Y,[X,Z]) = 0 \tag{A.106}$$

for any $Y,Z \in \mathfrak{p}$ and $X \in \mathfrak{h}$; use this with $X = [T_a,T_b]$, $Y = T_a$, and T_b. This gives

$$R_{abab} = -g_H([[T_a,T_b],T_b],T_a), \tag{A.107}$$

which is (A.104). □

By H-invariance, in an orthonormal basis g_H must be the Euclidean or the Minkowski metric. In the former case, for $n = 2$, the orthonormal basis (u_1,u_2) of $T_H(G/H) \cong \mathbb{R}^2$ may be either taken to be (T_1,T_2) for any ρ, or, for $\rho = 1$ and hence $O(3)$, the geometrically more natural basis discussed in §A.3, i.e., $(J_2,-J_1)$ (for the other two cases (T_1,T_2) was also the natural basis). Either way, the Lie brackets (A.92) - (A.94) or (A.21), (A.32), and (A.26) give

$$R_{1212} = \rho. \tag{A.108}$$

By (4.47) this also gives the sectional curvature, so that, so far in the Riemannian case,

$$k = \rho. \tag{A.109}$$

Eq. (A.104) is also valid in the Lorentzian case, but here one must be more careful about the choice of the basis (u_0,u_1) in $2d$ Minkowski space $(\mathbb{R}^2,\eta)$, with $\eta = \mathrm{diag}(-1,1)$.

We now study the three cases separately.

[663] See Kobayashi & Nomizu (1969), chapter XI, Theorems 3.2 and 3.3.

- As explained in §A.3, for $\rho = 1$, i.e., $G = O(1,2)$, we take $u_0 = -e_2$ and $u_1 = -e_1$. Eqs. (A.104) and (A.24) with $\eta(e_2,e_2) = \eta_{00} = -1$ then give $R_{0101} = -1$. Since the sectional curvature picks up a minus sign because of the denominator in (4.47), which in the Riemannian case equals $+1$ in an orthonormal basis but in the Lorentzian case equals -1, this gives $k = 1$.

- Similarly, for $\rho = -1$ and hence $G = O(2,1)$, with basis $u_0 = -f_2$ and $u_0 = -f_1$ of Minkowski $\mathbb{R}^2$, eqs. (A.104) and (A.28) give $R_{0101} = 1$ and hence $k = -1$.

- Finally, for $\rho = 0$ we obtain $R_{0101} = 0$ because in (4.12) the commutator vanishes:

$$[X,Y] = [T_1,T_2] = 0. \tag{A.110}$$

Hence we have $k = \rho$ in all six cases.

This proves Corollary 4.11 for $n = 2$. As a bridge to the general case $n \geq 2$, we note that

$$[u,v]w = \rho(\langle u,w\rangle v - \langle v,w\rangle u), \tag{A.111}$$

where the inner product is either the Euclidean or the Minkowski one, as the case requires. This follows from linear extension of (A.89) and hence has been derived for $n = 2$ only. But (A.111) holds in any dimension! To see this, we note that the adjoint action of $H = O(n)$ or $H = O(1,n-1)$ on $\mathfrak{g}$ consists of Lie algebra automorphisms (as this is true for all of G). Hence

$$[ku,kv] = \mathrm{Ad}(k)([u,v]) = k[u,v]k^{-1}, \tag{A.112}$$

for any $k \in H$ and $u,v \in \mathfrak{p} \cong \mathbb{R}^n$, with $[u,v] \in \mathfrak{so}(\mathfrak{n})$. If $n > 2$, we may take three mutually orthogonal vectors u,v,w and take k to be the reflection in the (hyper)plane orthogonal to w. Then by construction we have

$$ku = u; \qquad kv = v; \qquad k^{-1}w = kw = -w, \tag{A.113}$$

so that (A.112) gives

$$[u,v]w = -k([u,v]w). \tag{A.114}$$

By definition of k (which implies that $kx = -x$ is only true if x is a multiple of w), this implies that $[u,v]w$ is a multiple of w, which is impossible unless $[u,v]w = 0$. Therefore, $[u,v]$ maps any vector orthogonal to u and v to zero, which yields (A.111) for any n. The covariance property (A.112) has not only delivered the conclusion just given, but it also implies that the constant ρ in (A.111) is independent of the u-v plane (since H can move any plane to any other plane).

Given (A.111), the Lie algebra $\mathfrak{g}$ is now entirely known.[664] What remains is to find the right basis of $\mathfrak{g}$ for the three cases $\rho = 1,0,-1$, and thus recover the Lie algebras of

$$O(n+1); \qquad E(n); \qquad O(n,1) \tag{A.115}$$

in the Euclidean case, and in the Minkowski case, of

$$O(1,n); \qquad P(n); \qquad O(2,n-1). \tag{A.116}$$

Once again using (A.104) to show that $k = \rho$, this finishes the proof of Corollary 4.11. □

[664] The next step is best done in a basis provided by the root space decomposition of semi-simple Lie algebras, which requires more background than this appendix offers. Helgason (1978) is a complete reference for this.

B Background from formal PDE theory

This appendix collects some background for the study of the (gauged) Einstein equations as quasi-linear second-order hyperbolic PDEs. This field is huge and we just describe what we need for chapter 7. To start, all modern (i.e. post 1945) PDE theory is based on ***distributions***.

B.1 Distributions and Sobolev spaces on manifolds

This section collects some basic fact, more or less in *staccato* style, and without proofs.[665]

1. ***Notation.*** Let $n > 0$ and $x \in \mathbb{R}^n$. It is convenient to write $x = (x_1, \ldots, x_n)$ rather than our usual $(x^1, \ldots, x^n)$. Let $\alpha = (\alpha_1, \ldots \alpha_n)$, with $\alpha_i \in \mathbb{N}$ (where $0 \in \mathbb{N}$). We abbreviate

$$|\alpha| := \sum_{i=1}^{n} \alpha_i; \tag{B.1}$$

$$D^\alpha := \left(\frac{\partial}{\partial x_1}\right)^{\alpha_1} \cdots \left(\frac{\partial}{\partial x_n}\right)^{\alpha_n} \equiv \partial_1^{\alpha_1} \cdots \partial_n^{\alpha_n} \equiv \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \cdots \partial x_n^{\alpha_n}}; \tag{B.2}$$

$$x^\alpha := x_1^{\alpha_1} \cdots x_n^{\alpha_n}. \tag{B.3}$$

2. ***Test functions.*** For each measurable (usually open) subset $\Omega \subset \mathbb{R}^n$, let $\mathscr{D}(\Omega)$ be $C_c^\infty(\Omega)$ as a set, equipped with the topology in which $f_\lambda \to f$ iff there is a compact set $K \subset \Omega$ such that $\mathrm{supp}(f_\lambda) \subset K$ for all λ, and for all multi-indices α one has

$$\|D^\alpha(f_\lambda - f)\|_\infty \to 0. \tag{B.4}$$

   This *implies* $\mathrm{supp}(f) \subset K$ also for the limit function. This may be generalized to manifolds M, as follows. For some given atlas (U_i, φ_i) we say that $f_\lambda \to f$ in $\mathscr{D}(M) = C_c^\infty(M)$ iff for each $\psi_i \in C_c^\infty(U_i)$ and all multi-indices α one has

$$\|D^\alpha(\psi_i(f_\lambda - f) \circ \varphi_i^{-1})\|_\infty \to 0. \tag{B.5}$$

   This turns out to be independent of the choice of the atlas. Elements of $\mathscr{D}(\mathbb{R}^n)$, $\mathscr{D}(\Omega)$, or $\mathscr{D}(M)$ are all called ***test functions***.

   A ***rapidly decreasing (test) function*** $f \in \mathscr{S}(\mathbb{R}^n)$ is a function $f \in C^\infty(\mathbb{R}^n)$ for which the function $x \mapsto x^\alpha D^\beta f$ is bounded for all multi-indices α and β. One often writes

$$\langle x \rangle := (1 + \|x\|^2)^{1/2}, \tag{B.6}$$

   and uses $x \mapsto \langle x \rangle^\alpha D^\beta f$, which of course gives the same space. The topology on $\mathscr{S}(\mathbb{R}^n)$ is such that $f_\lambda \to f$ iff for all $l, m \in \mathbb{N}$ and multi-indices α and β with $|\alpha| \leq l$ and $|\beta| \leq m$,

$$\|x^\alpha D^\beta (f_\lambda - f)\|_\infty \to 0. \tag{B.7}$$

3. ***Distributions*** on Ω are elements of the space $\mathscr{D}'(\Omega)$ of all continuous maps $u : \mathscr{D}(\Omega) \to \mathbb{C}$. A linear map $u : \mathscr{D}(\Omega) \to \mathbb{C}$ is continuous in the topology just defined iff for each compact $K \subset \Omega$ there is $m \in \mathbb{N}$ and $C > 0$ such that for all α with $|\alpha| \leq m$,

$$|\langle u, f \rangle| \equiv |u(f)| \leq C \|D^\alpha f\|_\infty. \tag{B.8}$$

[665] For details see for example Hörmander (1990), §6.3, Taylor (1996), §4.3, and Grubb (2009), *passim*.

For example, a distribution of order zero (i.e. $m = 0$ is just a (Radon) measure on Ω.

The space $\mathscr{D}'(\Omega)$ carries the weak topology, in which $u_\lambda \to u$ iff $\langle u_\lambda, f\rangle \to \langle u, f\rangle$ for each $f \in \mathscr{D}(\Omega)$. In this topology, $\mathscr{D}(\Omega)$ is dense in $\mathscr{D}'(\Omega)$, where $u \in \mathscr{D}(\Omega)$ defines $u \in \mathscr{D}'(\Omega)$ through the L^2 inner product, i.e., $\langle u, f\rangle = \langle \overline{u}, f\rangle_{L^2(\Omega)}$. Adding a middle man gives a ***Gelfand triple***, in which each embedding is continuous and dense:

$$\mathscr{D}(\Omega) \subset L^2(\Omega) \subset \mathscr{D}'(\Omega). \tag{B.9}$$

Likewise for $\mathscr{D}(M)$, provided we equip our manifold M with a measure that in coordinates has the same null sets as Lebesgue measure.[666] For example, any (background) Riemannian metric on M provides such a measure (7.10). Also in that case we obtain a Gelfand triple

$$\mathscr{D}(M) \subset L^2(M) \subset \mathscr{D}'(M). \tag{B.10}$$

Tempered distributions on $\mathbb{R}^n$ are continuous linear maps $u : \mathscr{S}(\mathbb{R}^n) \to \mathbb{C}$. The (weak) topology on the ensuing space $\mathscr{S}'(\mathbb{R}^n)$ defines convergence $u_\lambda \to u$ of nets iff there are $l, m \in \mathbb{N}$ and $C > 0$ such that for all α with $|\alpha| \leq l$ and β with $|\beta| \leq m$ one has

$$|\langle u, f\rangle| \leq C\|x^\alpha D^\beta f\|_\infty. \tag{B.11}$$

Similarly to (B.9), one has a Gelfand triple (i.e. the embeddings are continuous and dense)

$$\mathscr{S}(\mathbb{R}^n) \subset L^2(\mathbb{R}^n) \subset \mathscr{S}'(\mathbb{R}^n), \tag{B.12}$$

and since $\mathscr{D}(\mathbb{R}^n) \subset \mathscr{S}(\mathbb{R}^n)$ continuously, and hence $\mathscr{S}'(\mathbb{R}^n) \subset \mathscr{D}'(\mathbb{R}^n)$, this extends to

$$\mathscr{D}(\mathbb{R}^n) \subset \mathscr{S}(\mathbb{R}^n) \subset L^2(\mathbb{R}^n) \subset \mathscr{S}'(\mathbb{R}^n) \subset \mathscr{D}'(\mathbb{R}^n). \tag{B.13}$$

4. ***Weak derivatives***. It will be convenient from now on to write, whenever convenient, $\langle u, f\rangle$ for $u(f)$. For each multi-index α, the ***weak derivative*** $D^\alpha u$ of $u \in \mathscr{D}'(\mathbb{R}^n)$ is defined by

$$\langle D^\alpha u, f\rangle := (-1)^{|\alpha|}\langle u, D^\alpha f\rangle. \tag{B.14}$$

This definition comes from the fake formula $\langle u, f\rangle = \int_{\mathbb{R}^n} d^n x\, u(x) f(x)$, which on repeated partial integration would give (B.14). Any linear partial differential operator may therefore be regarded as a map $L : \mathscr{D}'(\mathbb{R}^n) \to \mathscr{D}'(\mathbb{R}^n)$, with adjoint $L^* : \mathscr{D}(\mathbb{R}^n) \to \mathscr{D}(\mathbb{R}^n)$, i.e.,

$$\langle Lu, f\rangle = \langle u, L^* f\rangle. \tag{B.15}$$

For example, if $L = D^\alpha$, then $L^* = (-1)^{|\alpha|} D^\alpha$. The derivatives in Lu are called ***weak***, those in $L^* f$ being ***classical***. Similarly, a solution $u \in \mathscr{D}'(\mathbb{R}^n)$ of a linear PDE $Lu = F$ (with initial conditions), i.e. $\langle Lu, f\rangle = \langle u, L^* f\rangle$ for all $f \in \mathscr{D}(\mathbb{R}^n)$, is called ***weak***.

The definition (B.14) also applies to $u \in \mathscr{D}'(\mathbb{R}^n)$, at least if Ω is open in $\mathbb{R}^n$,[667] as well as to $\mathscr{D}'(M)$, provided M has no boundary (which indeed is our standing assumption).

[666] Hörmander's definition of a distribution on M coincides with the one above if we choose such a measure.

[667] Be careful with (B.15) if Ω is not open. For example, if $\Omega = [0,\infty) \times \mathbb{R}^n$ and $L = -\Box = \partial_t^2 - \Delta$, then (due to boundary terms in partial integration) the inhomogeneous wave equation $Lu = F$ with initial conditions $u(0,x) = f$ and $\dot{u}(0,x) = g(x)$ becomes $-\int_0^\infty dt \int_{\mathbb{R}^n} d^n x\, u \Box f = \int_0^\infty dt \int_{\mathbb{R}^n} d^n x F f + \int_{\mathbb{R}^n} d^n x\, g(x) f(0,x) - f(x)\dot{f}(0,x)$.

5. ***Sobolev spaces.*** For any $s \in \mathbb{N}$, based on (B.9), define the ***Sobolev space***

$$H^s(\Omega) := \{u \in L^2(\Omega) \mid D^\alpha u \in L^2(\Omega)\, \forall \alpha : |\alpha| \leq s\}, \tag{B.16}$$

where accordingly the derivatives inherent in D^α are weak. Clearly, $H^0(\Omega) = L^2(\Omega)$, but it can be shown that all $H^s(\Omega)$ are Hilbert spaces with respect to the inner product

$$\langle u, v\rangle_s := \sum_{|\alpha|\leq s} \langle D^\alpha u, D^\alpha v\rangle, \tag{B.17}$$

where $\sum_{|\alpha|\leq s}$ means $\sum_{\alpha:|\alpha|\leq s}$, and $\langle\cdot,\cdot\rangle$ is the inner product in $L^2(\Omega)$. Note the danger of ambiguous notation here: $\langle\cdot,\cdot\rangle_p$ often denotes the inner product in L^p, but here $\langle\cdot,\cdot\rangle_s$ stands for the inner product in H^s; in our notation the inner product in L^2 would be $\langle\cdot,\cdot\rangle_0$.

For $\Omega = \mathbb{R}^n$ a different perspective on Sobolev spaces comes from the ***Fourier transform***

$$\hat{f}(\xi) := (2\pi)^{-n/2} \int_{\mathbb{R}^n} d^n x\, f(x) e^{-i\xi x}; \tag{B.18}$$

$$\check{f}(x) := (2\pi)^{-n/2} \int_{\mathbb{R}^n} d^n \xi\, f(\xi) e^{i\xi x}, \tag{B.19}$$

which make sense as Lebesgue integrals for $f \in L^1(\mathbb{R}^n)$. If one also has $\hat{f} \in L^1(\mathbb{R}^n)$, then

$$\check{\hat{f}} = f. \tag{B.20}$$

The scope of these formulae may be extended in at least three different ways:[668]

(a) Eq. (B.18) yields a unitary isomorphism $L^2(\mathbb{R}^n) \xrightarrow{\cong} L^2(\mathbb{R}^n)$ of Hilbert spaces.

(b) The Fourier transform also defines a linear homeomorphism $\mathscr{S}(\mathbb{R}^n) \xrightarrow{\cong} \mathscr{S}(\mathbb{R}^n)$.

(c) Defining $\hat{f}$ for $f \in \mathscr{S}'(\mathbb{R}^n)$ by $\langle \hat{f}, f\rangle = \langle f, \check{f}\rangle$, the Fourier transform (B.18) even defines a linear homeomorphism $\mathscr{S}'(\mathbb{R}^n) \xrightarrow{\cong} \mathscr{S}'(\mathbb{R}^n)$ of tempered distributions.

Returning to Sobolev spaces, for $\Omega = \mathbb{R}^n$ may now (re)define, *for any* $s \in \mathbb{R}$,

$$H^s(\mathbb{R}^n) := \{u \in \mathscr{S}'(\mathbb{R}^n) \mid \xi \mapsto \langle\xi\rangle^s \hat{u}(\xi) \in L^2(\mathbb{R}^n)\}, \tag{B.21}$$

with inner product

$$\langle u, v\rangle_s := \int_{\mathbb{R}^n} d^n\xi\, \langle\xi\rangle^{2s} \overline{\hat{u}}(\xi)\hat{v}(\xi) = \int_{\mathbb{R}^n} d^n\xi\, (1 + \|\xi\|^2)^s\, \overline{\hat{u}}(\xi)\hat{v}(\xi) \tag{B.22}$$

For $s \in \mathbb{N}$ this reproduces (B.16) as a vector space (a fact that is not obvious), but the inner products (B.17) and (B.22) are different. Although they induce equivalent norms, for $s \in \mathbb{N}$ one has to specify which one is used. Either way, we have:

---

[668] If one equips $C_c^\infty(\mathbb{R}^n)$ with the unusual norm $\|f\|_0 = \max\{\|f\|_\infty, \|\hat{f}\|_\infty\}$, with associated completion denoted by $C_0^*(\mathbb{R}^n)$, then (B.18) yields an isometric isomorphism $C_0^*(\mathbb{R}^n) \xrightarrow{\cong} C_0^*(\mathbb{R}^n)$ as Banach spaces. For C*-algebra experts we note that the Fourier transform also yields an isomorphism $C^*(\mathbb{R}^n) \xrightarrow{\cong} C_0(\mathbb{R}^n)$ of commutative C*-algebras (here $C^*(\mathbb{R}^n)$ is the completion of $C_c^\infty(\mathbb{R}^n)$ in the operator norm obtained by letting $f \in C_c^\infty(\mathbb{R}^n)$ act on $L^2(\mathbb{R}^n)$ by convolution, whereas $C_0(\mathbb{R}^n)$ carries the supremum-norm). In this case (which follows from the Riemann–Lebesgue lemma) the Fourier transform is a special case of the Gelfand transform. See Landsman (2017), §C.15.

Theorem B.1 1. **Sobolev embedding theorem:** *For $m \geq 0$ and $s > m + \frac{1}{2}n$, one has*

$$H^s(\mathbb{R}^n) \subset C_b^m(\mathbb{R}^n), \tag{B.23}$$

where the embedding is continuous with respect to the norm $\|u\|_{m,\infty} = \sum_{|\alpha|\leq m} \|D^\alpha u\|_\infty$.

2. **Sobolev duality theorem:** *For any $s \in \mathbb{R}$ one has*

$$H^s(\mathbb{R}^n)^* \cong H^{-s}(\mathbb{R}^n), \tag{B.24}$$

i.e. $\Lambda \in H^s(\mathbb{R}^n)^*$ linearly, bijectively, and isometrically corresponds to $f \in H^{-s}(\mathbb{R}^n)$ via

$$\Lambda(u) = \int_{\mathbb{R}^n} d^n x\, f(x)u(x) \equiv \langle f, u\rangle. \tag{B.25}$$

3. *For $s > 0$ we have our third Gelfand triple*

$$H^s(\mathbb{R}^n) \subset L^2(\mathbb{R}^n) \subset H^{-s}(\mathbb{R}^n), \tag{B.26}$$

which analogously to (B.13) may be extended to a "Gelfand quintuple"

$$\mathscr{S}(\mathbb{R}^n) \subset H^s(\mathbb{R}^n) \subset L^2(\mathbb{R}^n) \subset H^{-s}(\mathbb{R}^n) \subset \mathscr{S}'(\mathbb{R}^n). \tag{B.27}$$

6. Sobolev spaces can also be defined on manifolds. For $u \in \mathscr{D}'(M)$, we define $u \in H^2(M)$ iff for each chart (U_i, φ_i) and $\chi_i \in C_c^\infty(V_i)$, where $V_i = \varphi_i(U_i) \subset \mathbb{R}^n$, the distribution $u \circ \varphi_i^{-1}\chi_i$ on $\mathscr{D}(\mathbb{R}^n)$, defined on $f \in \mathscr{D}(\mathbb{R}^n)$ by $\langle u \circ \varphi_i^{-1}\chi_i, f\rangle = \langle u, (\chi_i f) \circ \varphi_i\rangle$, is in $H^s(\mathbb{R}^n)$.

Theorem B.2 *Let M be a compact Riemannian manifold.*

1. *For each $s \in \mathbb{R}$ the space $\mathscr{D}(M)$ is dense in $H^s(M)$.*
2. *For each $s \in \mathbb{R}$ we have an isometric (Banach space) isomorphism*

$$H^s(M)^* \cong H^{-s}(M), \tag{B.28}$$

understood in the following way:[669] any continuous functional $\Lambda \in H^s(M)^*$ corresponds linearly, bijectively, and isometrically to $f \in H^{-s}(M)$ via

$$\Lambda(u) = \langle \overline{f}, u\rangle_{L^2(M)}. \tag{B.29}$$

3. **Sobolev embedding theorem:** *If $s > \frac{1}{2}n + k$, then $H^s(M) \subset C_b^k(M)$, where the embedding is continuous with respect to the norm $\|u\|_{m,\infty} = \sum_{|\alpha|\leq m} \|D^\alpha u\|_\infty$ on $C^k(M)$.*
4. **Rellich theorem:** *For $s \in \mathbb{R}$ and $\delta > 0$, the injection $H^{s+\delta}(M) \hookrightarrow H^s(M)$ is compact.*
5. *For $s > 0$ we have our final Gelfand triple, cf. (B.26),*

$$H^s(M) \subset L^2(M) \subset H^{-s}(M). \tag{B.30}$$

B.2 Linear wave equations

For the PDEs of interest to GR, $\mathbb{R}^n$ will be space, and time needs to be treated separately. Typically, for fixed time $T>0$ one considers Banach spaces like $C([0,T],H^s(\mathbb{R}^n))$, with norm

$$\|u\|_\infty = \sup_{t\in[0,T]} \|u(t)\|_s, \tag{B.31}$$

or $C^1([0,T],H^s(\mathbb{R}^n))$ with analogous norm, or $L^p([0,T],H^s(\mathbb{R}^n))$, $1\leq p<\infty$, normed by

$$\|u\|_p = \left(\int_0^T dt\,(\|u(t)\|_s)^p\right)^{1/p}, \tag{B.32}$$

or $L^\infty([0,T],H^s(\mathbb{R}^n))$, with norm

$$\|u\|_\infty = \operatorname{ess\,sup}_{t\in[0,T]}\|u(t)\|_s. \tag{B.33}$$

Here we define $L^p([0,T],H^s(\mathbb{R}^n))$, $1\leq p<\infty$, as the completion of $C([0,T],H^s(\mathbb{R}^n))$ in the norm (B.32), and also (avoiding Banach space-valued measurable functions), *define* the space $L^\infty([0,T],H^s(\mathbb{R}^n))$ as the (Banach) dual of $L^1([0,T],H^{-s}(\mathbb{R}^n))$, in that we identify $f\in L^\infty([0,T],H^s(\mathbb{R}^n))$ with the functional $\Lambda_f\in(L^1([0,T],H^{-s}(\mathbb{R}^n)))^*$ given by, cf. (B.29),

$$\Lambda_f(g) = \int_0^T dt\,\langle f(t),g(t)\rangle. \tag{B.34}$$

To see such spaces in action, we consider the free wave equation on $\mathbb{R}^{n+1}$, i.e.

$$(-\partial_t^2+\Delta)u = F; \qquad u(0,x)=f; \qquad \dot u(0,x)=g(x). \tag{B.35}$$

For $F=0$ and $n=1,3$, the (unique) solution (known since the 18th century) is

$$u(t,x) = \tfrac{1}{2}\left(f(x+t)-f(x-t)+\int_{x-t}^{x+t} dy\,g(y)\right); \qquad (n=1); \tag{B.36}$$

$$u(t,x) = \frac{1}{4\pi t^2}\int_{|y-x|=t} d\sigma^2(y)\left(tg(y)+f(y)-\sum_{i=1}^3 \partial_i f(y)(x_i-y_i)\right); \qquad (n=3). \tag{B.37}$$

From this, we see that in $n=1$ the solution at (t,x) only depends on initial data within its causal past $J^-(x,t)$, intersected with the Cauchy surface $\Sigma=\{(x^0=0,x),x\in\mathbb{R}^n\}$. Indeed, recall the causal past $J^-(t,x)$, emanating from (t,x), and its boundary $E^-(t,x)$, i.e. the past lightcone,

$$J^-(t,x) = \{(y^0,y)\in\mathbb{R}^{n+1}, |y^0-x^0|\geq|y-x|, y^0\leq x^0\}; \tag{B.38}$$

$$E^-(t,x) = \{(y^0,y)\in\mathbb{R}^{n+1}, |y^0-x^0|=|y-x|, y^0\leq x^0\}, \tag{B.39}$$

cf. (5.90) - (5.91) with $y^0\geq x^0$ replaced by $y^0\leq x^0$ (as well as x by (t,x), etc.). In $n=1$,

$$\Sigma\cap J^-(x,t) = \{(y^0=0,y), y\in[x-t,x+t]\}. \tag{B.40}$$

In $n=3$ the solution $u(t,x)$ even depends on the initial data at $\Sigma\cap E^-(x,t)$ only, since

$$\Sigma\cap E^-(t,x) = \{(y^0=0,y), |y-x|=t\}. \tag{B.41}$$

[669] Also, $H^s(M)^*\cong H^s(M)$ through its own inner product; the pairing in (B.25) is through the L^2 inner product.

An analogous phenomenon holds in the inhomogeneous case $F \neq 0$, in which case the solution

$$u(t,x) = \frac{1}{4\pi} \int_{E^-(t,x)} d^3\sigma(s,y) \, \frac{F(s,y)}{|(s-t,y-x)|}, \tag{B.42}$$

for zero initial data for simplicity, only depends on the values of F at the past lightcone $E^-(t,x)$. In other words, $F(s,y)$ only influences u along the forward lightcone emanating from (s,y). The situation in $n = 3$ (and also in all higher *odd* spatial dimensions), in which both initial data f, g and the inhomogeneous term F affect the solution only along future light rays is called the ***strong Huygens principle***. The (ordinary) ***Huygens principle***, then, formalizes the situation in $n = 1, 2$, and all higher *even* dimensions, in which the entire causal future of (s,y) affects the solution–or, equivalently, $u(t,x)$ only depends on data within its causal past.

An explicit solution for any F, f, and g may be written down using the Fourier transform:

$$\hat{u}(t,\xi) = \cos(t|\xi|)\hat{f}(\xi) + \frac{\sin(t|\xi|)}{|\xi|}\hat{g}(\xi) + \int_0^t ds \, \frac{\sin((t-s)|\xi|)}{|\xi|}\hat{F}(s,\xi); \tag{B.43}$$

as the notation indicates, the formula (B.18) is only applied to the x-variable, and, within the function classes to be discussed, the actual solution $u(t,x)$ may be (re)constructed from (B.19). Although the space-time and causal structure of the solution is not at all obvious from this formula, the advantage is that (B.43) easily implies an ***energy inequality***: for any $s \in \mathbb{Z}$,

$$\|u(t,\cdot)\|_{s+1} + \|\dot{u}(t,\cdot)\|_s \leq C_{s,T} \left((\|f\|_{s+1} + \|g\|_s) + \int_0^T d\tau \, \|F(\tau,\cdot)\|_s \right), \tag{B.44}$$

where $0 < T < \infty$, provided that $F \in L^1([0,T], H^s(\mathbb{R}^n))$, $f \in H^{s+1}(\mathbb{R}^n)$, and $g \in H^s(\mathbb{R}^n)$, so that the right-hand side makes sense. The proof is an exercise, using the fact that (B.22) implies

$$\|u(t)\|_s^2 = \int_{\mathbb{R}^n} d^n\xi \, (1 + \|\xi\|^2)^s \, |\hat{u}(t,\xi)|^2. \tag{B.45}$$

Corollary B.3 *For any $T > 0$ and $s \in \mathbb{Z}$, the free wave equation (B.35) with initial conditions $f \in H^{s+1}(\mathbb{R}^n)$ and $g \in H^s(\mathbb{R}^n)$, and $F \in L^1([0,T], H^s(\mathbb{R}^n))$, has a unique solution*

$$u(t,x) \in C([0,T], H^{s+1}(\mathbb{R}^n)) \cap C^1([0,T], H^s(\mathbb{R}^n)). \tag{B.46}$$

Uniqueness follows either from the derivation of the explicit solution (B.43) from the initial data, or from (B.44): if u_1 and u_2 both solve (B.35), then $u = u_1 - u_2$ solves (B.35) for $F = f = g = 0$, so that the right-hand side and hence the left-hand side of (B.44) vanishes, etc.

We now turn to linear wave equations of the form $Lu = F$ with initial data (B.35), and

$$L = g^{\rho\sigma}(t,x)\partial_\rho\partial_\sigma + b^\rho(t,x)p_\rho + a(t,x). \tag{B.47}$$

Since we don't have an explicit solution, the derivation of a suitable energy inequality (to be used as a lemma for proving existence, uniqueness, and analytic properties of solutions) will have to be *a priori*.[670] A particularly useful energy inequality for the operator (B.47) is

$$\sum_{|\alpha|\leq 1} \|D^\alpha u(t,\cdot)\|_s \leq C'_{s,T} \left(\sum_{|\alpha|\leq 1} \|D^\alpha u(0,\cdot)\|_s + \int_0^t d\tau \, \|Lu(\tau,\cdot)\|_s \right). \tag{B.48}$$

[670] These *a priori* derivations are straightforward but very lengthy, and therefore we simply state the results without derivation; for (B.48) see Sogge (2008), §I.3 and Luk (undated), §4. See also Ringström (2009) for similar estimates.

This inequality is valid for any $0 < t < T < \infty$, $s \in \mathbb{Z}$, and u such that (B.46) holds,[671] with $Lu \in L^1([0,T], H^s)$. It immediately gives uniqueness by the same argument as for the free wave equation, but existence and regularity require a more advanced, functional-analytic argument.[672]

In order to explain the reasoning, let us first take a simpler situation. For $\Omega \subset \mathbb{R}^n$, let

$$L : \mathscr{D}'(\Omega) \to \mathscr{D}'(\Omega) \tag{B.49}$$

be a linear operator, e.g. as in (B.47), with adjoint $L^* : \mathscr{D}(\Omega) \to \mathscr{D}(\Omega)$ defined by (B.15). As already mentioned, the PDE $Lu = F$ (with zero initial conditions for simplicity) then means

$$\langle u, L^* f \rangle = \langle F, f \rangle \tag{B.50}$$

for all $f \in \mathscr{D}(\Omega)$. Throughout the argument, we must assume that, for any net (f_λ) in $\mathscr{D}(\Omega)$,

$$L^* f_\lambda \to L^* f \;\Rightarrow\; f_\lambda \to f. \tag{B.51}$$

If L^* is a bijection, and $F \in \mathscr{D}'(\Omega)$, which is the very least regularity to impose, then we are done at the coarsest level of proving existence and uniqueness of a solution $u \in \mathscr{D}'(\Omega)$, since its value at $\psi \in \mathscr{D}'(\Omega)$ is given by finding the unique $f \in \mathscr{D}(\Omega)$ for which $\psi = L^* f$, and putting

$$\langle u, \psi \rangle = \langle F, f \rangle. \tag{B.52}$$

The assumption (B.51) then implies that if $\psi_\lambda \to \psi$, i.e., $L^* f_\lambda \to L^* f$, then $f_\lambda \to f$, and hence $\langle F, f_\lambda \rangle \to \langle F, f \rangle$ since $F \in \mathscr{D}'(\Omega)$ by assumption, and hence $\langle u, \psi_\lambda \rangle \to \langle u, \psi \rangle$, since $\langle u, \psi_\lambda \rangle = \langle F, f_\lambda \rangle$. Thus u is a continuous linear functional on $\mathscr{D}(\Omega)$ and hence $u \in \mathscr{D}'(\Omega)$.

If L^*, still assumed to be injective, merely has dense range $\operatorname{ran}(L^*) \subset \mathscr{D}(\Omega)$, then one still has existence and uniqueness of u, since for $\psi \in \operatorname{ran}(L^*)$ eq. (B.52) continues to apply, whereas for ψ outside the range of L^* we may write $\psi = \lim_\lambda L^* f_\lambda$ and then $\langle u, \psi \rangle = \lim_\lambda \langle F, f_\lambda \rangle$.

Finally, if L^*, still injective, does not have dense range, the Hahn–Banach theorem (for locally convex vector spaces) yields existence of u by extending the solution $u : \operatorname{ran}(L^*) \to \mathbb{C}$ constructed above to a continuous linear map $u : \mathscr{D}'(\Omega) \to \mathbb{C}$, but one loses uniqueness. Fortunately, in many applications to PDEs uniqueness still follows from suitable energy inequalities.

Such inequalities also play a central role in refining the above argument. Suppose one has two Gelfand(ish) triples $\mathscr{D}(\Omega) \subset W \subset \mathscr{D}'(\Omega)$ and $\mathscr{D}(\Omega) \subset Z \subset \mathscr{D}'(\Omega)$, where W and Z are Banach spaces and all inclusion maps are continuous with dense image, and suppose that

$$\|f\|_Z \leq C \|L^* f\|_W \quad (\forall f \in \mathscr{D}(\Omega)). \tag{B.53}$$

This 'energy condition' supersedes the continuity assumption (B.51) within $\mathscr{D}(\Omega)$, and is also more powerful in that it clearly implies that L is injective, which is an essential condition for the whole analysis to apply in the first place. Furthermore, the inequality (B.53) implies:

Provided L^* is injective, for any $F \in Z^*$ there is a solution $u \in W^*$ to $Lu = F$.

Note that $\mathscr{D}(\Omega) \subset Z$ implies $Z^* \subset \mathscr{D}'(\Omega)$, and similarly $\mathscr{D}(\Omega) \subset W$ implies $W^* \subset \mathscr{D}'(\Omega)$. Compared with the earlier argument where the assumption $F \in \mathscr{D}'(\Omega)$ gave a solution $u \in \mathscr{D}'(\Omega)$,

[671] Moreover, the derivation requires that $g^{\mu\nu}(t,x)$, $b^\mu(t,x)$, and $a(t,x)$ be C^∞ with uniform bounds on all derivatives, where $(t,x) \in [0,T] \times \mathbb{R}^n$, as well as $\sum_{\mu,nu} |g^{\mu\nu}(t,x) - \eta^{\mu\nu}| \leq \frac{1}{2}$, where η is the Minkowski metric.

[672] The following arguments are adapted from Vasy (2015), chapter 17. The entire book is very useful.

we have now strengthened the assumption to $F \in Z^* \subset \mathscr{D}'(\Omega)$, and, given (B.53), accordingly strengthened the conclusion $u \in \mathscr{D}'(\Omega)$ to $u \in W^* \subset \mathscr{D}'(\Omega)$. Indeed, noting that

$$\mathrm{ran}(L^*) \subset \mathscr{D}(\Omega) \subset W, \tag{B.54}$$

let $\psi \in \mathrm{ran}(L^*)$, so $\psi = L^* f$, and define a linear map $u : W \to \mathbb{C}$ initially on $\mathrm{ran}(L^*) \subset W$ by

$$\langle u, L^* f \rangle_{W^*-W} = \langle F, f \rangle_{Z^*-Z}. \tag{B.55}$$

Because of (B.53), if $L^* f_\lambda \to L^* f$ in W, then $f_\lambda \to f$ in Z, and hence on the assumption $F \in Z^*$, the functional u defined by (B.55) is continuous on $\mathrm{ran}(L^*)$ in the (norm) topology of W. Once again, the Hahn–Banach extension theorem (but this time simply for Banach spaces) gives a continuous extension $u : W \to \mathbb{C}$, i.e. $u \in W^*$, as claimed.

We now show how the energy estimate (B.48) implies an estimate à la (B.53). For any $T > 0$, we replace u in (B.48) by $f \in C_c^\infty((0,T) \times \mathbb{R}^n)$, which certainly satisfies the assumptions validating (B.48), and replace L by L^*. Then $D^\alpha u(0,\cdot)$ is replaced by $D^\alpha f(0,\cdot) = 0$. Furthermore, for any multi-index α, $s \in \mathbb{R}$, $k \in \mathbb{N}$, and $f \in H^s$, by definition of the Sobolev spaces we have

$$\|f\|_{-s} \leq C' \sum_{|\alpha| \leq k} \|D^\alpha f\|_{-s-k}. \tag{B.56}$$

With $k = 1$, also using the trivial estimate $\int_0^t d\tau\, g(\tau) \leq \int_0^T d\tau\, g(\tau)$ for $0 < t < T$ and $g(\tau) \geq 0$, in this case with $g(\tau) = \|L^* f(\tau,\cdot)\|_{-s-1}$, we find, for any $s \in \mathbb{Z}$ and $f \in C_c^\infty((0,\infty) \times \mathbb{R}^n)$,

$$\|f(t,\cdot)\|_{-s} \leq C \int_0^T d\tau\, \|L^* f(\tau,\cdot)\|_{-s-1}. \tag{B.57}$$

This is a special case of (B.53), with

$$W = L^1([0,T], H^{-s-1}(\mathbb{R}^n)); \qquad Z = C([0,T], H^{-s}(\mathbb{R}^n)); \tag{B.58}$$

$$W^* = L^\infty([0,T], H^{s+1}(\mathbb{R}^n)); \qquad Z^* \supset L^1([0,T], H^s(\mathbb{R}^n)). \tag{B.59}$$

The precise form of Z^* (which is the space of bounded measures on $[0,T]$ taking values in H^s) is not needed here. Assuming zero initial conditions for the moment, the abstract argument above gives a solution $u \in L^\infty([0,T], H^{s+1}(\mathbb{R}^n))$ for $F \in L^1([0,T], H^s(\mathbb{R}^n))$, which, by the original energy inequality (B.48) is also unique. More advanced arguments involving elliptic regularity further push the solution into (B.46).[673] Finally, the case of nonzero initial data f, g can be reduced to the case $f = g = 0$ by a standard trick. For given F, let v solve $Lv = F$ for zero initial data. Define $w(t,x) = f(x) + tg(x)$. Then $u = v + w$ solves $Lu = F$ for given f, g. Thus:

Theorem B.4 *For any $T > 0$, let L be defined by (B.47), including all assumption stated afterwards. For any $s \in \mathbb{Z}$, the linear wave equation $Lu = F$, with $F \in L^1([0,T], H^s(\mathbb{R}^n))$ and initial conditions $f \in H^{s+1}(\mathbb{R}^n)$ and $g \in H^s(\mathbb{R}^n)$, see (B.35), has a unique solution*

$$u(t,x) \in C([0,T], H^{s+1}(\mathbb{R}^n)) \cap C^1([0,T], H^s(\mathbb{R}^n)). \tag{B.60}$$

The Sobolev embedding theorem (B.23) then pushes this into the smooth realm:

Corollary B.5 *In the setting of the previous theorem, if F, f, and g are smooth, then so is u.*

One can also show that the causal properties of the solution relative to F and the initial data f, g are the same as for the free wave equation, except that the *strong* Huygens principle need not apply. But the 'ordinary' one, implying causal propagation of initial data and F, always does.

[673] See Sogge (2008), p. 20.

B.3 Quasi-linear wave equations

In either the (naive) wave gauge or its refinement the $\hat{g}$-wave gauge, the vacuum Einstein equations (7.121) have the abstract form $Lu = F$, where $u = g_{\mu\nu}$ and L is like (B.47), with the difference that in $L = g^{\rho\sigma}(u)\partial_\rho\partial_\sigma$ the coefficient of the highest (i.e. second) order derivative now depends on u, and furthermore $F = F(u,\partial u)$ depends on u and ∂u. Such equations (in the more general case that g and F may depend on u, ∂u, and even (t,x)) are called ***quasi-linear***, and if the signature of g is Lorentzian, as we of course assume, the PDE is ***hyperbolic***.[674]

We assume for the moment that u takes values in $\mathbb{R}$; the generalization to $u = (g_{\mu\nu})$ taking values in $\mathbb{R}^{10}$, is straightforward and will be outlined shortly. It is also sufficient for basic applications to GR to assume that $g^{\rho\sigma} : \mathbb{R} \to \mathbb{R}$ is smooth, as is $F : \mathbb{R}\times\mathbb{R}^{n+1} \to \mathbb{R}$. So we study

$$g^{\rho\sigma}(u)\partial_\rho\partial_\sigma u = F(u,\partial u). \tag{B.61}$$

As opposed to truly nonlinear hyperbolic PDEs, the quasi-linear case is relatively easy because it can be solved by reduction to the linear case. One can only feel fortunate that the Einstein equations (at least in a suitable gauge) fall into this category. Here is the basic result:[675]

Theorem B.6 *Let F be smooth and $g^{\rho\sigma}$ smooth and not too far from the Minkowski metric.*[676]

1. *With $f := u(0,\cdot) \in H^{s+1}(\mathbb{R}^n)$ and $g := \dot{u}(0,\cdot) \in H^s(\mathbb{R}^n)$, eq. (B.61) has a unique solution*

$$u \in L^\infty([0,T],H^{s+1}(\mathbb{R}^n)); \qquad \dot{u} \in L^\infty([0,T],H^s(\mathbb{R}^n)), \tag{B.62}$$

   provided $s > \frac{1}{2}n$. Here T is either arbitrary (as in the linear case),[677] *or there exists*

$$T_* = T_*(\|f\|_{s+1},\|g\|_s) \tag{B.63}$$

   such that $\|D^\alpha u\|_\infty = \infty$ on $[0,T_*)\times\mathbb{R}^n$, for some $|\alpha| \le 2$.

2. *This u depends continuously on the initial data, i.e. if $f_k \to f$ in $H^{s+1}(\mathbb{R}^n)$ and $g_k \to g$ in $H^s(\mathbb{R}^n)$, then $u_k \to u$ in $L^\infty([0,T],H^{s+1}(\mathbb{R}^n))$ with $\dot{u}_k \to \dot{u}$ in $L^\infty([0,T],H^s(\mathbb{R}^n))$.*

3. *If $f \in C_c^\infty(\mathbb{R}^n)$ and $g \in C_c^\infty(\mathbb{R}^n)$, then $u \in C^\infty([0,T)\times\mathbb{R}^n)$, cf. Corollary B.5.*

Eq. (B.61) is solved using a generalization of the Picard iteration procedure.[678] Take

$$u_0(x) = f(x) = u(0,x), \tag{B.64}$$

and iteratively define u_{k+1} as the solution to the inhomogeneous linear PDE

$$g^{\rho\sigma}(u_k)\partial_\rho\partial_\sigma u_{k+1} = F(u_k,\partial u_k), \tag{B.65}$$

[674] In fluid mechanics all these dependencies also occur, see e.g. Taylor (1996), chapter 16.

[675] See Sogge (2008), §I.4, Luk (undated), §6, Choquet-Bruhat (2009), App. III, or Ringström (2009), chapter 9.

[676] Think of $\sum_{\rho,\sigma}\|g^{\rho\sigma} - \eta^{\rho\sigma}\|_\infty \le \frac{1}{2}$, as in Sogge (2008). For initial data $f \in H^{s+1}(\mathbb{R}^n)$ and $g \in H^s(\mathbb{R}^n)$, one can make further (contrived) regularity assumptions on $g^{\rho\sigma}$ and F that push u into (B.60). See Ringström, Ch. 9.

[677] For an example with $T_* < \infty$, take $(\partial_t^2 - \Delta)u = u^3$ with $u(0,x) = \dot{u}(0,x) = 1$ (times a cutoff function), so that $u(t,x) = 1/(1-t)$ (for small x), and hence $T^* = 1$.

[678] Recall that an ODE $u'(t) = f(t,u(t))$ with initial condition $u(0) = u_0$, which is equivalent to the integral equation $u(t) = u_0 + \int_0^t ds\, f(s,u(s))$, may be solved by iteration from $u_0(t) = u_0$ and $u_{k+1}(t) = u_0 + \int_0^t ds\, f(s,u_k(s))$. For suitably regular f, this sequence (u_k) uniformly converges to a solution u on some interval $[0,T]$.

subject to the initial conditions $u_{k+1}(0,x) = f(x)$ and $\dot{u}_{k+1}(0,x) = g(x)$, as for u itself.[679] For given $u_k(t,x)$, eq. (B.65) is the type of PDE studied in the previous section. Hence Theorem B.4 guarantees a solution for any $T > 0$, but convergence of the iteration and uniformity of the energy inequality (B.44) in k, gives a less regular solution in Theorem B.6 compared to the linear case.

Theorem B.6 applies to the Einstein equations in the ($\hat{g}$-) wave gauge, except that:

- Instead of a single unknown u we now have 10 unknowns $g_{\mu\nu}$, with one equation for each (but the ensuing system is coupled, since $g^{\rho\sigma}$ is a function of all $g_{\mu\nu}$ and so is $F(g,\partial g)$).
- The Cauchy surface $\{t = 0\} \subset \mathbb{R}^{n+1}$ is replaced by a $3d$ (Riemannian) manifold Σ.
- The initial data $u(0,\cdot) = f$ and $\dot{u}(0,\cdot) = g$ are replaced by the Cauchy data $(\tilde{g},\tilde{k}$ on Σ.
- Using either local coordinate patches and a partition of unity, or a background metric $\hat{\gamma}$ on Σ making the construction coordinate-independent (like the $\hat{g}$-wave gauge), one can define Sobolev spaces $H^s(\Sigma)$ for any $s \in \mathbb{R}$ (in view of $s < \frac{1}{2}n + 1$ in Theorem B.6, $s \in \mathbb{N}$ is enough).[680] This construction may be extended from functions on Σ to arbitrary tensors $\tau \in \mathfrak{X}^{(k,l)}(\Sigma)$, yielding Sobolev spaces $H^s_{(k,l)}(\Sigma)$. Thus one may say, e.g., $\tilde{k} \in H^s_{(2,0)}(\Sigma)$.
- The PDE (B.65) is replaced by the reduced (vacuum) Einstein equations (7.121).

This eventually leads to Theorem 7.16 in §7.6 and its localization Proposition 7.17. Much as uniqueness is proved from an energy inequality, the localized uniqueness of the above kind is proved from a localized energy inequality. We merely explain this for the free wave equation $\Box u = 0$ in $\mathbb{R}^{n+1}$, but the principle is the same also in Lorentzian geometry.[681]

For any $0 \leq t \leq R$, $(t,x) \in \mathbb{R}^{n+1}$, and (reasonable) function $u(t,x)$, define

$$E(t,x,R) = \tfrac{1}{2}\int_{|y-x|\leq R-t} d^n y\,[\dot{u}(t,y)^2 + \nabla u(t,y)\cdot\nabla u(t,y)]. \tag{B.66}$$

This is just the energy of u, restricted to the ball $B(x;R-t) \subset \mathbb{R}^n$. If $\Box u = 0$, then

$$0 \leq s \leq t \;\;\Rightarrow\;\; 0 \leq E(t,x,R) \leq E(s,x,R). \tag{B.67}$$

That is, $t \mapsto E(t,x,R)$ is monotonically non-increasing. Fix $R > 0$, and note that

$$E(0,x,R) = \tfrac{1}{2}\int_{B(x,R)} d^n y(g(y)^2 + \nabla f(y)\cdot\nabla f(y)). \tag{B.68}$$

Eq. (B.68) implies that if $f(y) = g(y) = 0$ for all y such that $|y-x| \leq R$, then $E(0,x,R) = 0$, and hence $E(t,x,R) = 0$ for all $0 \leq t \leq R$ by (B.67), and hence $u(t,x) = 0$ by (B.68). Taking $R = t$ shows that if $f(y) = g(y) = 0$ for all y such that $|y-x| \leq t$, then $u(t,x) = 0$. In other words, if $f = g = 0$ within $\Sigma_0 \subset \Sigma$ (defined as the $t = 0$ hyperplane $\mathbb{R}^n_0$ in $\mathbb{R}^{n+1}$), then $u = 0$ within $D^+(\Sigma)$. Equivalently, if $u_1 = u_2$ and $\dot{u}_1 = \dot{u}_2$ at Σ_0, then $u_1 = u_2$ in $D^+(\Sigma_0)$. In case of the Einstein equations, $u_1 = u_2$ becomes $g_1 \cong g_2$ (isometrically), but otherwise the reasoning is similar, ultimately based on the property $g_1 = g_2$ if both metrics are brought into the same gauge.

[679] This works if $f, g \in C^\infty_c(\mathbb{R}^n)$. For initial data $f \in H^s(\mathbb{R}^n)$ and $g \in H^{s+1}(\mathbb{R}^n)$ one needs to approximate f and g within the spaces mentioned by sequences (f_k) and (g_k) in $C^\infty_c(\mathbb{R}^n)$, respectively, upon which the initial conditions for (B.65) change into $u_{k+1}(0,x) = f_{k+1}(x)$ and $\dot{u}_{k+1}(0,x) = g_{k+1}(x)$.

[680] See Taylor (1996), Vol. I, chapter 4, Ringström (2009), chapter 15, or Choquet-Bruhat (2009), Appendix I.

[681] See Choquet-Bruhat (2009), Appendix III, Theorem 2.15.

Literature

History of general relativity:[682] Primary sources

Alexandrow, W. (1923). Über den kugelsymmetrischen Vakuumvorgang in der Einsteinschen Gravitationstheorie. *Annalen der Physik* 377, 141–152.

Arnowitt, R., Deser, S., Misner, C.W. (1962). The dynamics of general relativity. *Gravitation: An Introduction to Current Research*, ed. Witten, L., pp. 227–264 (Wiley). Reprinted in *General Relativity and Gravitation* 40, 1997–2027 (2008) and also available as `gr-qc/0405109`.

Avez, A. (1963). Essais de géométrie riemannienne hyperbolique globale. Applications à la relativité générale, *Annales de l'institut Fourier* 132, 105–190.

Bergmann, P.G. (1949). Non-linear field theories. *Physical Review* 75, 680–685.

Bergmann, P.G. (1958). Conservation laws in general relativity as the generators of coordinate transformations. *Physical Review* 112, 287–289.

Bergmann, P.G. (1961). Observables in general relativity. *Reviews of Modern Physics* 33, 510–514.

Bergmann, P.G., Brunings, J.H.M. (1949). Non-linear field theories II. Canonical equations and quantization. *Reviews of Modern Physics* 21, 480–487.

Bergson, H. (1922). *Durée et Simultanéité. A Propos de la Théorie d'Einstein* (Alcan).

Birkhoff, G.D. (1923). *Relativity and Modern Physics* (Harvard University Press).

Bott, R., Mather, J.M. (1968). Topics in topology and differential geometry. *Batelle Rencontres: 1967 Lectures in Mathematics and Physics*, eds. DeWitt, C., Wheeler, J.A., pp. 460–515 (W.A. Benjamin).

Boyer, R.H., Lindquist, R.W. (1967). Maximal analytic extension of the Kerr metric. *Journal of Mathematical Physics* 8, 265–281.

Cassirer, E. (1921). *Zur Einstein'schen Relativitätstheorie* (Bruno Cassirer).

Choquet-Bruhat, Y. (1967). Hyperbolic partial differential equations on a manifold. *Batelle Rencontres: 1967 Lectures in Mathematics and Physics*, eds. DeWitt, C., Wheeler, J.A., pp. 84–106 (W.A. Benjamin).

Choquet-Bruhat, Y., Geroch, R. (1969). Global aspects of the Cauchy problem in general relativity. *Communications in Mathematical Physics* 14, 329–335.

Darmois, G. (1927). Les équations de la gravitation einsteinienne. *Mémorial des sciences mathématiques*, fascicule 25. `http://www.numdam.org/article/MSM_1927__25__1_0.pdf`.

DeWitt, B.S. (1967). Quantum theory of gravity. I. The canonical theory. *Physical Review* 160, 1113–1148.

Dirac, P.A.M. (1950). Generalized Hamiltonian systems. *Canadian Journal of Mathematics* 12, 129–148.

Dirac, P.A.M. (1958a). Generalized Hamiltonian dynamics. *Proceedings of the Royal Society of London A* 246, 326–332.

Dirac, P.A.M. (1958b). The theory of gravitation in Hamiltonian form. *Proceedings of the Royal Society of London A* 246, 333–343.

Droste, J. (1916). *Het zwaartekrachtsveld van een of meer lichamen volgens de theorie van Einstein*. PhD Thesis, Leiden University.

Droste, J. (1917). The field of a single centre in Einstein's theory of gravitation, and the motion of a particle in that field. *Proceedings Royal Netherlands Academy of Arts and Sciences (Amsterdam)* 19, 197–215.

Eddington, A.S. (1923). *The Mathematical Theory of Relativity* (Cambridge University Press).

Eddington, A.S. (1924). A comparison of Whitehead's and Einstein's formulae. *Nature* 113, 192.

Einstein, A. (1905). Zur Elektrodynamik bewegter Körper, *Annalen der Physik* 17, 891–921.

Einstein, A. (1915a). Zur allgemeinen Relativitätstheorie, *Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften (Berlin)* 778–786.

[682] The cut is 1970. In order to find a reference cited in the main text, the reader should guess into which of the four categories it falls (or check all of them): History of general relativity: Primary sources; History of general relativity: Secondary sources; Books; Articles and online resources. The advantages outweigh the disadvantages!

Einstein, A. (1915b). Zur allgemeinen Relativitätstheorie (Nachtrag). *Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften (Berlin)* 799–801.

Einstein, A. (1915c). Erklärung der Perihelbewegung des Merkur aus der allgemeinen Relativitätstheorie, *Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften (Berlin)* 831–839.

Einstein, A. (1915d). Die Feldgleichungen der Gravitation, *Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften (Berlin)* 844–847.

Einstein, A. (1916a). Die Grundlage der allgemeinen Relativitätstheorie, *Annalen der Physik* 49, 769-822.

Einstein, A. (1916b). Hamiltonsches Prinzip und allgemeine Relativitätstheorie, *Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften (Berlin)* 1111–1116.

Einstein, A. (1917a). Über die spezielle und allgemeine Relativitätstheorie (Braunschweig).

Einstein, A. (1917b). Kosmologische Betrachtungen zur allgemeinen Relativitätstheorie, *Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften (Berlin)* 142–152.

Einstein, A. (1918a). Prinzipielles zur allgemeinen Relativitätstheorie, *Annalen der Physik* 55, 241–244.

Einstein, A. (1918b). Besprechung: Weyl, Hermann, Raum - Zeit - Materie. Vorlesungen über Allgemeine Relativitätstheorie. *Die Naturwissenschaften* 25, 373.

Einstein, A. (1918c). Über Gravitationswellen. *Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften (Berlin)* 154–167.

Einstein, A. (1918d). Kritisches zu einer von Hrn. de Sitter gegebenen Lösung der Gravitationsgleichungen. *Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften (Berlin)* 448–472.

Einstein, A. (1920). Über die spezielle und die allgemeine Relativitätstheorie (Vieweg). English translation:

Einstein, A. (1921). *Relativity: The Special and General Theory* (Henry Holt and Company).

Einstein, A. (1934). On the method of theoretical physics (The Herbert Spencer lecture, delivered at Oxford, June 10, 1933). *Philosophy of Science* 1, 163–169.

Einstein, A. (1939). On a stationary system with spherical symmetry consisting of many gravitating masses *Annals of Mathematics* 40, 922–936.

Einstein, A. (1996a). *The Collected Papers of Albert Einstein, Vol. 4: The Swiss Years: Writings, 1912–1914*, eds. Klein, M.J., Kox, A.J., Renn, J., Schulman, R. (Princeton University Press). `https://einsteinpapers.press.princeton.edu/vol4-doc/`.

Einstein, A. (1996b). *The Collected Papers of Albert Einstein, Vol. 6: The Berlin Years: Writings, 1914–1917*, eds. Klein, M.J., Kox, A.J., Schulman, R. (Princeton University Press). `https://einsteinpapers.press.princeton.edu/vol6-doc/`.

Einstein, A. (1999). *The Collected Papers of Albert Einstein, Vol. 8, Part A: The Berlin Years: Correspondence 1914-1917*, eds. Schulmann, R., Kox, A.J., Janssen, M., Illy, J. (Princeton University Press). `https://einsteinpapers.press.princeton.edu/vol8a-doc/`.

Einstein, A. (2002a). *The Collected Papers of Albert Einstein, Vol. 7: The Berlin Years: Writings, 1918–1921*, eds. Janssen, M. *et al* (Princeton University Press). `https://einsteinpapers.press.princeton.edu/vol7-doc/`.

Einstein, A. (2002b). *The Collected Papers of Albert Einstein, Vol. 7: The Berlin Years: Writings, 1918–1921. (English translation of selected texts)*, Translated by Alfred Engel (Princeton University Press). `https://einsteinpapers.press.princeton.edu/vol7-trans/`.

Einstein, A. (2013). *The Collected Papers of Albert Einstein, Vol. 13: The Berlin Years: Writings & Correspondence January 1922–March 1923*, eds. Buchwald, D.K., Illy, J., Rosenkranz, Z., Sauer, T. (Princeton University Press). `https://einsteinpapers.press.princeton.edu/vol13-doc/`.

Einstein, A., Grommer, J. (1927). *Allgemeine Relativitätstheorie und Bewegungsgesetz* (Verlag der Akademie der Wissenschaften).

Einstein, A., Grossmann, M. (1913). *Entwurf einer verallgemeinerten Relativitätstheorie* (Teubner).

Einstein, A., Infeld, L., Hoffman, B. (1938). The gravitational equations and the problem of motion. *Annals of Mathematics* 39, 65–100.

Einstein, A., Lorentz, H.A., Minkowski, H., Weyl, H. (1923). *The Principle of Relativity: A Collection of Original Memoirs on the Special and General Theory of Relativity, Notes by A. Sommerfeld* (Methuen and Company).

Einstein, A., Rosen, N. (1935). The particle problem in the general theory of relativity. *Physical Review* 48, 73–77.

Eisland, J. (1925). The group of motions of an Einstein space. *Transactions of the American Mathematical Society* 27, 231–245.

Fermi, E. (1922). Sopra i fenomeni che avvengono in vicinanza di una linea oraria. *Atti della Reale Accademia Nazionale dei Lincei, Classe di Scienze Fisiche, Matematiche e Naturali* 31, 184–187, 306–309.

Finkelstein, D. (1958). Past-future asymmetry of the gravitational field of a point particle. *Physical Review* 110, 956–967.

Fourès-Bruhat, Y. (1948). Sur l'intégration des équations d'Einstein. *Comptes Rendus de l'Académie des Sciences* 226, 48–51.

Fourès-Bruhat, Y. (1952). Théorème d'existence pour certains systèmes d'équations aux dérivées partielles non linéaires. *Acta Mathematica* 88, 141–225.

Fourès-Bruhat, Y. (1956). Sur l'intégration des équations de la relativité générale. *Journal of Rational Mechanics and Analysis* 5, 951–966.

Gauss, C.F. (1828). *Disquisitiones generales circa superfies curvas* (Typis Dieterichianis). English translation: *General Investigations of Curved Surfaces*, `http://www.gutenberg.org/files/36856/36856-pdf.pdf`.

Geroch, R.P. (1966). Singularities in closed universes. *Physical Review Letters* 17, 445–447.

Geroch, R. (1968). What is a singularity in General Relativity? *Annals of Physics (N.Y.)* 48, 526–540.

Geroch, R. (1970). Domain of dependence, *Journal of Mathematical Physics* 11, 437–449.

Gödel, K. (1949). An example of a new type of cosmological solutions of Einstein's field equations of gravitation. *Reviews of Modern Physics* 21, 447–450.

Hadamard, J. (1923). *Lectures on Cauchy's problem in linear partial differential equations* (Yale University Press).

Hawking, S.W. (1965). *Properties of Expanding Universes*. PhD Thesis, University of Cambridge. `https://www.repository.cam.ac.uk/handle/1810/251038`.

Hawking, S.W. (1966). Singularities and the geometry of spacetime. Adams Prize Essay. Reprinted in *European Journal of Physics H* 39, 413–503 (2014).

Hawking, S.W. (1967). The occurrence of singularities in cosmology. III. Causality and singularities. *Proceedings of the Royal Society (London)* A300, 187–201.

Hawking, S.W., Penrose, R. (1970). The singularities of gravitational collapse and cosmology. *Proceedings of the Royal Society (London)* A314, 529–548.

Hilbert, D. (1900). Über den Zahlbegriff. *Jahresbericht des deutschen Mathematiker Vereinigung* 8, 180–184.

Hilbert, D. (1901). Über Flächen von constanter Gaussscher Krümmung. *Transactions of the American Mathematical Society* 2, 87–99.

Hilbert, D. (1902a). Mathematical Problems. Lecture delivered before the International Congress of Mathematicians at Paris in 1900. *Bulletin of the American Mathematical Society* 8, 437–479. Translated from *Göttinger Nachrichten*, 1900, pp. 253–297.

Hilbert, D. (1902b). Über die Grundlagen der Geometrie. *Mathematische Annalen* 56, 381–422.

Hilbert, D. (1904). Über das Dirichletsche Prinzip. *Mathematische Annalen* 59, 161–186. *Nachrichten von der Königlichen Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse*, 395–407.

Hilbert, D. (1915). Die Grundlagen der Physik (Erste Mitteilung). *Nachrichten von der Königlichen Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse*, 395–408.

Hilbert, D. (1917). Die Grundlagen der Physik (Zweite Mitteilung). *Nachrichten von der Königlichen Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse*, 53–76.

Hilbert, D. (1992). *Natur und mathematisches Erkennen: Vorlesungen, gehalten in 1919–1920 in Göttingen*, ed. Rowe, D. (Springer).

Hoffmann, B. (1932a). *On the Spherically Symmetric Field in Relativity*. PhD Thesis, Harvard University.

Hoffmann, B. (1932b). On the spherically symmetric field in relativity. *The Quarterly Journal of Mathematics* 3, 226–237.

Israel, W. (1967). Event horizons in static vacuum space-times. *Physical Review* 164, 1776–1779.

Israel, W. (1968). Event horizons in static electrovac space-times. *Communications in Mathematical Physics* 8, 245–260.

Jebsen, J.T. (1921). Über die allgemeinen kugelsymmetrischen Lösungen der Einsteinschen Gravitationsgleichungen im Vakuum. *Arkiv för Matematik, Astronomi och Fysik*, 15, 1–9. Translated and reprinted as: On the general spherically symmetric solutions of Einstein's gravitational equations in vacuo. *General Relativity and Gravitation* 37, 2253–2259 (2005).

Kerr, R.P. (1963). Gravitational field of a spinning mass as an example of algebraically special metrics. *Physical Review Letters* 11, 237–238.

Kretschmann, E. (1917). Über den physikalischen Sinn der Relativitätspostulate: A. Einsteins neue und seine ursprüngliche Relativitätstheorie. *Annalen der Physik* 53, 575–614.

Kronheimer, E.H., Penrose, R. (1967). On the structure of causal spaces. *Mathematical Proceedings of the Cambridge Philosophical Society* 63, 481–501.

Lemaître, G. (1933). L'univers en expansion. *Annales de la Société scientifique de Bruxelles* A 53, 51–85. Translated and reprinted as The expanding Universe, *General Relativity and Gravitation* 29, 641–680 (1997), with editorial note and biography by Andrzej Krasiński, pp. 637–40.

Leray, J. (1953). *Hyperbolic Differential Equations* (Mimeographed Lecture Notes, The Institute for Advanced Study).

Levi-Civita, T. (1917a). Nozione di parallelismo in una varietà qualunque e conseguente specificazione geometrica della curvatura riemanniana. *Rendiconti del Circolo Matematico di Palermo* XLII, 173–215.

Levi-Civita, T. (1917b). Realtà fisica di alcuni spazî normali del Bianchi, *Rendiconti della Reale Accademia dei Lincei* 26, 519–31.

Levi-Civita, T. (1926). *The Absolute Differential Calculus* (Blackie & Son).

Lichnerowicz, A. (1939). *Sur Certains Problèmes Globaux Relatifs au Système des Équations d'Einstein*. PhD Thesis, Université de Paris. `http://archive.numdam.org/article/THESE_1939__226__1_0.pdf`.

Lichnerowicz, A. (1944). L'intégration des équations de la gravitation relativiste et le problème des n-corps. *Journal de Mathématiques Pures et Appliquées* 23, 37–63.

Lichnerowicz, A. (1955). *Théories Relativistes de la Gravitation et de l'Électromagnetisme* (Masson).

Lorentz, H.A. (1916). Over Einstein's theorie der zwaartekracht, I-IV. *Koninklijke Akademie van Wetenschappen te Amsterdam. Verslagen van de Gewone Vergaderingen der Wis- en Natuurkundige Afdeeling* 24, 1389–1402, 1759–1774, 25, 468–486, 1380–1396. English translation in Lorentz, H.A. (1937), *Collected Papers, Vol. 5*, pp. 246–313 (Nijhoff).

Manasse, F.K., Misner, C.W. (1963). Fermi normal coordinates and some basic concepts in differential geometry. *Journal of Mathematical Physics* 4, 735–745.

Misner, C.W. (1963). The flatter regions of Newman, Unti, and Tamburino's generalized Schwarzschild space. *Journal of Mathematical Physics* 4, 924–937.

Noether, E. (1918). Invariante Variationsprobleme. *Nachrichten von der Königlichen Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse* 253–260.

Nordström, G. (1918). On the energy of the gravitational field in Einstein's theory. *Koninklijke Akademie van Wetenschappen te Amsterdam. Verslagen van de Gewone Vergaderingen der Wis- en Natuurkundige Afdeeling* 26, 1201–1208.

Oppenheimer, J.R., Volkoff, G.M. (1939). On massive neutron cores. *Physical Review* 55, 374–381.

Oppenheimer, J.R., Snyder, H. (1939). On continued gravitational contraction. *Physical Review* 56, 455–459.

Pauli, W. (1921). Relativitätstheorie. *Encyklopädie der mathematischen Wissenschafen* Vol. V19 (Teubner). Translation: *Theory of Relativity* (Pergamon Press, 1958, Dover, 1981).

Penrose, R. (1963). Null hypersurface initial data for classical fields of arbitrary spin and for general relativity. *Aerospace Research Laboratories* 63–65. Reprinted in *General Relativity and Gravitation* 12, 225–264 (1980).

Penrose, R. (1964). Conformal treatment of infinity. *Relativity, Groups, and Topology*, eds. DeWitt, B., DeWitt-Morette, C.M. , pp. 565–584 (Gordon & Breach). Reprinted in *General Relativity and Gravitation* 43, 901–922 (2011).

Penrose, R. (1965). Gravitational collapse and space-time singularities. *Physical Review Letters* 14, 57–59.

Penrose, R. (1966). *An analysis of the structure of space-time*. Adams Prize Essay. *Roger Penrose: Collected Works, Volume 1: 1953–1967*, pp. 579–730 (Oxford University Press, 2011).

Penrose, R. (1968). *Structure of space-time*. *Batelle Rencontres: 1967 Lectures in Mathematics and Physics*, eds. DeWitt, C., Wheeler, J.A., pp. 121–235 (W.A. Benjamin).

Penrose, R. (1969). Gravitational collapse: The role of general relativity. *Rivista del Nuovo Cimento, Numero Speziale I*, 252. Reprinted in *General Relativity and Gravitation* 34, 1141–1165 (2002).

Poincaré, H. (1895). Analysis Situs. *Journal de l'École Polytechnique. Serié 11.* Translation (by J. Stillwell): *Papers on Topology: Analysis Situs and Its Five Supplements.* `https://www.maths.ed.ac.uk/~v1ranick/papers/poincare2009.pdf`

Racine, C. (1934). *Le problème des n corps dans la théorie de la Relativité*. PhD Thesis, University of Paris.

Reichenbach, H. (1924). *Axiomatik der relativistische Raum-Zeit-Lehre* (Vieweg).

Reichenbach, H. (1928). *Philosophie der Raum-Zeit-Lehre* (Walter de Gruyter).

Reissner, H. (1916). Über die Eigengravitation des elektrischen Feldes nach der Einsteinschen Theorie. *Annalen der Physik* 50, 106–120.

Ricci, M.M.G, Levi-Civita, T. (1901). Méthodes de calcul différentiel absolu et leurs applications. *Mathematische Annalen* 54, 125–201.

Riemann, B. (1854). Über die Hypothesen, welche der Geometrie zu Grunde liegen (Habilitationsvortrag). *Abhandlungen der Königlichen Gesellschaft der Wissenschaften zu Göttingen* 13, 132–152 (1867).

Riemann, B. (1876). *Gesammelte Werke und Wissenschaftlicher Nachlass*, eds. R. Dedekind & H. Weber (Teubner). `https://www.emis.de/classics/Riemann/Geom.pdf`.

Rindler, W. (1956). Visual horizons in world models. *Monthly Notices of the Royal Astronomical Society* 116, 662–677.

Robb, A.A. (1914). *A Theory of Time and Space* (Cambridge University Press).

Robb, A.A. (1936). *Geometry of Time and Space* (Cambridge University Press).

Schlick, M. (1922). *Raum und Zeit in der gegenwärtigen Physik* (Springer).

Schouten, J.A.. (1918). Die direkte Analysis zur neueren Relativitätsttheorie, *Verhandelingen der Koninklijke Akademie van Wetenschappen te Amsterdam*, 12, 1–95.

Schouten, J.A.. (1924). *Der Ricci-Kalkül* (Springer, 1924).

Schouten, J. A., Struik, D. J. (1936). *Einführung in die neueren Methoden der Differentialgeometrie* (Noordhoff).

Schrödinger, E. (1950). *Space-Time Structure* (Cambridge University Press).

Schwarzschild, K. (1916). Über das Gravitationsfeld eines Massenpunktes nach der einsteinschen Theorie. *Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften (Berlin)* 189–196. English translation: On the gravitational field of a mass point according to Einstein's theory. `arXiv:physics/9905030`.

Sitter, W. de (1917a). On the relativity of inertia: Remarks concerning Einstein's latest hypothesis. *Proceedings of the Royal Netherlands Academy of Arts and Sciences (KNAW)* 19, 1217–1225. `https://www.dwc.knaw.nl/DL/publications/PU00012455.pdf`.

Sitter, W. de (1917b). On the curvature of space. *Proceedings of the Royal Netherlands Academy of Arts and Sciences (KNAW)* 20: 229–243. `https://www.dwc.knaw.nl/DL/publications/PU00012216.pdf`.

Veblen, O., Whitehead, J.H.C. (1932). *The Foundations of Differential Geometry* (Cambridge University Press).

Weyl, H. (1913). *Die Idee der Riemannschen Fläche* (Teubner).

Weyl, H. (1917). Zur Gravitationstheorie. *Annalen der Physik* 54, 117–145.

Weyl, H. (1918a). *Raum - Zeit - Materie: Vorlesungen über Allgemeine Relativitätstheorie* (Springer). English translation (of the fourth edition from 1921): *Space - Time - Matter* (Methuen, 1922).

Weyl, H. (1918b). Reine Infinitesimalgeometrie. *Mathematische Zeitschrift* 2, 384–411.

Whitney, H. (1936). Differentiable manifolds. *Annals of Mathematics* 37, 645–680.

Zeeman, E.C. (1964). Causality implies the Lorentz group. *Journal of Mathematical Physics* 5, 490–493.

History of general relativity: Secondary sources

Antoci, S. (2003). David Hilbert and the origin of the "Schwarzschild solution". arXiv:physics/0310104.

Antoci, S., Liebscher, E. (2001). Reconsidering Schwarzschild's original solution. *Astronomical Notes* 322, 137–142. arXiv:gr-qc/0102084.

Ashtekar, A. (2014). The last 50 years of general relativity and gravitation: from GR3 to GR20 Warsaw conferences. *General Relativity and Gravitation* 46:1706.

Barbour, J.B. (1989). *Absolute or Relative Motion? Volume 1: The Discovery of Dynamics* (Cambridge University Press).

Biezunski, M. (1987). Einstein's reception in Paris in 1922. *The Comparative Reception of Relativity*, ed. Glick, T.F., pp. 169–188 (Springer).

Blum, A., Lalli, R., Renn, J. (2015). The reinvention of general relativity: A historiographical framework for assessing one hundred years of curved space-time. *Isis* 106, 598–620.

Blum, A.S., Lalli, R., Renn, J. (2016). The renaissance of general relativity: How and why it happened. *Annalen der Physik* 528, 344–349.

Blum, A.S., Lalli, R., Renn, J., eds. (2020). *The Renaissance of General Relativity in Context* (Springer).

Brezis, H., Browder, F. (1998). Partial differential equations in the 20th century. *Advances in Mathematics* 135, 76–144.

Carter, B. (2006). Half a century of black hole theory: From physicists' purgatory to mathematicians' paradise. *AIP Conference Proceedings* 841, 29–50.

Choquet-Bruhat, Y. (2014). Beginnings of the Cauchy problem. arXiv:1410.3490.

Choquet-Bruhat, Y. (2018). *A Lady Mathematician in this Strange Universe: Memoirs* (World Scientific).

Corry, L. (1999). From Mie's electromagnetic theory of matter to Hilbert's unified foundations of physics, *Studies in History and Philosophy of Modern Physics* 30, 159–183.

Corry, L. (2004). *David Hilbert and the Axiomatization of Physics (1898–1918): From Grundlagen der Geometrie to Grundlagen der Physik* (Kluwer).

Corry, L. (2018). Hilbert's sixth problem: Between the foundations of geometry and the axiomatization of physics. *Philosophical Transactions of the Royal Society A*. DOI: 10.1098/rsta.2017.0221.

Darrigol, O. (2014). The mystery of Riemann's curvature. *Historia Mathematica* 42, 47–83.

DeWitt-Morette, C. (2011). *The Pursuit of Quantum Gravity: Memoirs of Bryce DeWitt from 1946 to 2004* (Springer).

Dell'Aglio, L. (1997). On the genesis of the concept of covariant differentiation. *Revue d'histoire des mathématiques* 2, 215–264.

Dieks, D. (2006). Another look at general covariance and the equivalence of reference frames. *Studies in History and Philosophy of Modern Physics* 37, 174–191.

Dieks, D. (2018). Time, coordinates and clocks: Einstein's struggle, arXiv:1801.09297.

Dongen, J. van (2010). *Einstein's Unification* (Cambridge University Press).

Dongen, J. van (2017). The epistemic virtues of the virtuous theorist: On Albert Einstein and his autobiography. *Epistemic Virtues in the Sciences and the Humanities, Boston Studies in the Philosophy and History of Science, Vol. 321*, eds. Dongen, J. van, Paul, H., pp. 63–77 (Springer).

Earman, J. (1999). The Penrose–Hawking singularity theorems: History and implications. *The Expanding Worlds of General Relativity (Einstein Studies Vol. 7)*, eds. Goenner, H., Renn, J., Ritter, T., Sauer, T., pp. 236–267 (Birkhäuser).

Earman, J., Eisenstaedt, J. (1999). Einstein and singularities. *Studies in History and Philosophy of Modern Physics* 30, 185–235.

Eckes, C. (2019). Hermann Weyl in Göttingen (1904–1913): The combined impact of Hilbert, Klein and Zermelo, *Bhavana: The Mathematics Magazine* 3 (January 2019), https://bhavana.org.in/hermann-weyl-part2/.

Eggertsson, R. (2019). *The Noether Theorems*. MSc Thesis, Radboud University Nijmegen. Available on request.

Ehlers, J. (2007). A K Raychaudhuri and his equation. *Pramana–Journal of Physics* 69, 7–14.

Ehlers, J., Krasiński, A. (2006). Comment on the paper by J.T. Jebsen. *General Relativity and Gravitation* 38, 1329–1330.

Eisenstaedt, J. (1993). Lemaître and the Schwarzschild solution. *The Attraction of Gravitation: New Studies in the History of General Relativity*, Earman, J., Janssen, M., Norton, J.D. (eds.), pp. 353–389 (Birkhäuser).

Eisenstaedt, J. (2006). *The Curious History of Relativity: How Einstein's Theory of Gravity Was Lost and Foud Again* (Princeton University Press).

Ellis, G.F.R. (2014). Stephen Hawking's 1966 Adams Prize Essay. *European Journal of Physics H* 39, 403–411.

Farwell, R., Knee, C. (1990). The missing link: Riemann's "Commentatio,", differential geometry and tensor analysis. *Historia Mathematica* 17, 223–255.

Fokker, A.D. (1955). Albert Einstein: 14 maart 1878 (*sic*)–18 april 1955, *Nederlands Tijdschrift voor Natuurkunde*, May 1955, pp. 125–129.

Fölsing, A. (1993). *Albert Einstein: Eine Biographie* (Suhrkamp).

Freeman, K. (2008). *A Historical Overview of Connections in Geometry*, M.Sc. Thesis, Wichita State University, `http://www.math.wichita.edu/~pparker/classes/Freeman_Kamielle_SP2011.pdf`.

Friedrich, H. (2011). Editorial note to: Roger Penrose, Conformal treatment of infinity. *General Relativity and Gravitation* 43, 897–900.

Giovanelli, M. (2013). Erich Kretschmann as a proto-logical-empiricist: Adventures and misadventures of the point-coincidence argument. *Studies in History and Philosophy of Modern Physics* 44, 115–134.

Giovanelli, M. (2019). Nothing but coincidences: The point-coincidence argument and Einstein's struggle with the meaning of coordinates in Physics. `http://philsci-archive.pitt.edu/16830/`.

Giulini, D. (2007). Remarks on the notions of general covariance and background independence, *Lecture Notes in Physics* 721, 105–120.

Godart, O. (1992). Contributions of Lemaître to General Relativity (1922–1934). *Studies in the History of General Relativity*, eds. Eisenstaedt, J., Kox, A.J., pp. 437–452 (Birkhäuser).

Goenner, H.F.M. (2004). On the history of unified field theories. *Living Reviews in Relativity* 7, 2–129.

Goenner, H.F.M. (2017). A golden age of general relativity? Some remarks on the history of general relativity. *General Relativity and Gravitation* 49:42.

Goodstein, J.R. (2018). *Einstein's Italian Mathematicians: Ricci, Levi-Civita, and the Birth of General Relativity* (American Mathematical Society).

Gorban, A.N. (2018). Hilbert's sixth problem: The endless road to rigour. Introduction to a theme issue *Hilbert's' sixth problem* of the *Philosophical Transactions of the Royal Society A* 376, 2118.

Gray, J. (1999). *The Symbolic Universe: Geometry and Physics 1890–1930* (Oxford University Press).

Gray, J. (2007). *Worlds Out of Nothing: A Course in the History of Geometry in the 19th Century* (Springer).

Hawking, J. (1999). *Music to Move the Stars: A Life With Stephen* (MacMillan). Second edition:

Hawking, J. (2007). *Travelling to Infinity: My Life With Stephen* (Alma Books).

Heidegger, M. (1927). *Sein und Zeit* (Niemeyer).

Hodges, A. (2014). *Extra Time: Professor Sir Roger Penrose in conversation with Andrew Hodges*.
`https://www.youtube.com/watch?v=zN5eLsI_Tuo` (part 1);
`https://www.youtube.com/watch?v=FFWbpHmll1g` (part 2).

Hossenfelder, S. (2018). *Lost in Math: How Beauty Leads Physics Astray* (Basic Books).

Isaacson, W. (2017). *Einstein: His Life and Universe* (Simon & Schuster).

Israel, W. (1987). Dark stars: The evolution of an idea. *Three Hundred Years of Gravitation*, eds. Hawking, S.W., Israel, W., pp. 199–276. (Cambridge University Press).

Iurato, G. (2016). On the history of Levi-Civita's parallel transport. `arXiv:1608.04986`.

Janssen, M. (1992). H.A. Lorentz's attempt to give a coordinate-free formulation of the general theory of relativity. *Studies in the History of General Relativity*, eds. J. Eisenstaedt and A.J. Kox, pp. 344–361 (Birkhäuser).

Janssen, M. (2012). The twins and the bucket: How Einstein made gravity rather than motion relative in general relativity, *Studies in History and Philosophy of Modern Physics* 43, 159–175.

Janssen, M. (2014). "No success like failure . . . ": Einstein's quest for general relativity, 1907–1920. *The Cambridge Companion to Einstein*, eds. Janssen, M., Lehner, C., pp. 167–227 (Cambridge University Press).

Janssen, M., Lehner, C., eds. (2014). *The Cambridge Companion to Einstein* (Cambridge University Press).

Janssen, M., Renn, J. (2020). *How Einstein Found His Field Equations: A Source Book* (Springer).

Johansen, N.V., Ravndal, F. (2006). On the discovery of Birkhoff's theorem. *General Relativity and Gravitation* 38, 537–540.

Kaiser, D. (1998) A ψ is just a ψ? Pedagogy, practice, and the reconstitution of general relativity, 1942–1975. *Studies in History and Philosophy of Modern Physics* 29, 321–338.

Kossmann-Schwarzbach, Y. (2011). *The Noether Theorems: Invariance and Conservation Laws in the Twentieth Century* (Springer).

Kossmann-Schwarzbach, Y. (2020). The Noether theorems in context. `arXiv:2004.09254`.

Kox, A.J. (1988). Hendrik Antoon Lorentz, the ether, and the general theory of relativity. *Archive for History of Exact Sciences* 38, 67–78.

Kragh, H. (2007). *Conceptions of the Cosmos. From Myths to the Accelerating Universe: A History of Cosmology* (Oxford University Press).

Lalli, R. (2017). *Building the General Relativity and Gravitation Community During the Cold War* (Springer).

Landsman, K. (2021). Singularities, black holes, and cosmic censorship: A tribute to Roger Penrose. *Foundations of Physics* 51:42.

Landsman, K. (2022). Quantum theory and functional analysis. *Oxford Handbook of the History of Interpretations and Foundations of Quantum Mechanics*, ed. O. Freire, to appear (Oxford University Press).

Lehmkuhl, D. (2014). Why Einstein did not believe that general relativity geometrizes gravity, *Studies in History and Philosophy of Modern Physics* 46, 316–326.

Lehmkuhl, D. (2019). The Equivalence Principle(s). `http://philsci-archive.pitt.edu/17709/`. *Routledge Companion to the Philosophy of Physics*, eds. Knox, E., Wilson, A., to appear.

Lichnerowicz, A. (1992). Mathematics and general relativity: A recollection. *Studies in the History of General Relativity*, Eisenstaedt, J., Kox, A.J. (eds)., pp. 103–108 (Birkhäuser).

Lightman, A. (1989). *AIP Oral History Interviews: Roger Penrose*. `https://www.aip.org/history-programs/niels-bohr-library/oral-histories/34322`.

Longair, M.S. (2006). *The Cosmic Century: A History of Astrophysics and Csomology* (Cambridge University Press).

Melia, F. (2009). *Cracking the Einstein Code: Relativity and the Birth of Black Hole Physics* (University of Chicago Press).

Moore, G.H. (1995). The axiomatization of linear algebra: 1875–1940. *Historia Mathematica* 22, 262–303.

Nordmann, C. (1922). Einstein expose et discute sa théorie. *Revue des Deux Mondes* 9, 130–166.

Norton, J.D. (1985). What was Einstein's principle of equivalence? *Studies in History and Philosophy of Science Part A* 16, 203–246.

Norton, J.D. (1989). Coordinates and covariance: Einstein's view of space-time and the modern view. *Foundations of Physics* 19, 1215–1263.

Norton, J.D. (1992). The physical content of general covariance. *Studies in the History of General Relativity*, eds. Eisenstaedt, J., Kox, A.J., pp. 281–315 (Birkhäuser).

Norton, J.D. (1993). General covariance and the foundations of general relativity: Eight decades of dispute. *Reports on Progress in Physics* 56, 791–858.

Norton, J.D. (1995). Did Einstein stumble? The debate over general covariance. *Erkenntnis* 42, 223–245.

Norton, J.D. (1999). Geometries in collision: Einstein, Klein, and Riemann. *The Symbolic Universe*, ed. Gray, J., pp. 128–144 (Oxford University Press).

Norton, J.D. (2000). 'Nature is the realisation of the simplest conceivable mathematical ideas': Einstein and the canon of mathematical simplicity. *Studies in History and Philosophy of Modern Physics* 31, 135–170.

Norton, J.D. (2018). Einstein's conflicting heuristics: The discovery of general relativity. `http://philsci-archive.pitt.edu/14965/`.

Norton, J.D. (2019). The Hole Argument. *The Stanford Encyclopedia of Philosophy (Summer 2019 Edition)*, ed. Zalta, E.N. https://plato.stanford.edu/archives/sum2019/entries/spacetime-holearg/.

Nussbaumer, H., Bieri, L. (2009). *Discovering the Expanding Universe* (Cambridge University Press).

Pais, A. (1982). *Subtle is the Lord: The Science and Life of Albert Einstein* (Oxford University Press).

Penrose, R. (2015). *The Art of the Impossible: MC Escher and Me - Secret Knowledge*. Documentary. https://www.youtube.com/watch?v=f7kW8xd8p4s&t=15s and https://www.youtube.com/watch?v=1CYrGpd8k5w.

Read, J., Teh, N., Roberts, B., eds. (2021) *The Philosophy and Physics of Noether's Theorems*. (Cambridge University Press).

Reich, K. (1973). Die Geschichte der Differentialgeometrie van Gauß bis Riemann (1828–1868). *Archive for the History of Exact Sciences* 11, 273–376.

Reich, K. (1994). *Die Entwicklung des Tensorkalküls: Vom absoluten Differentialkalkül zur Relativitätstheorie* (Birkhäuser).

Reid, C. (1970). *Hilbert* (Springer).

Renn, J., ed. (2007). *The Genesis of General Relativity. Volumes 1-4* (Springer).

Renn, J., Gutfreund, H. (2015). *The Road to Relativity: The History and Meaning of Einstein's 'The Foundation of General Relativity'* (Princeton University Press).

Renn, J., Stachel, J. (2007). Hilbert's Foundation of Physics: From a theory of everything to a constituent of general relativity. *The Genesis of General Relativity. Volume 4*, ed. Renn, J., pp. 857–973 (Springer).

Ringström, H. (2010). Cosmic censorship for Gowdy spacetimes. *Living Reviews in Relativity* 13:2.

Ringström, H. (2015). Origins and development of the Cauchy problem in general relativity. *Classical and Quantum Gravity* 32:124003.

Robinson, D.C. (2009). Four decades of black hole uniqueness theorems. *The Kerr Spacetime: Rotating Black Holes in General Relativity*, eds. Wiltshire, D.L., Visser, M., Scott, S.M., pp. 115–143 (Cambridge University Press).

Robinson, D.C. (2019). Gravitation and general relativity at King's College London/ *The European Physical Journal H* 44, 181–270.

Rowe, D.E. (2006). Review of *"zwei wirkliche Kerle": Neues zur Entdeckung der Gravitationsgleichungen der Allgemeinen Relativitätstheorie durch Albert Einstein und David Hilbert* by Daniela Wuensch, *Historia Mathematica* 33, 491–508.

Rowe, D.E. (2018). *A Richer Picture of Mathematics: The Göttingen Tradition and Beyond* (Springer).

Rowe, D.E. (2021). *Emmy Noether: Mathematician Extraordinaire* (Springer).

Ryckman, T. (2005). *The Reign of Relativity: Philosophy in Physics 1915–1925* (Oxford University Press)

Ryckman, T. (2018). Early philosophical interpretations of general relativity. *The Stanford Encyclopedia of Philosophy (Spring 2089 Edition)*, ed. Zalta, E.N. https://plato.stanford.edu/archives/spr2018/entries/genrel-early/.

Rynasiewicz, R. (2014). Newton's Views on Space, Time, and Motion. *The Stanford Encyclopedia of Philosophy (Summer 2014 Edition)*, ed. Zalta, E.N. https://plato.stanford.edu/archives/sum2014/entries/newton-stm/.

Salisbury, J.D. (2020). Toward a quantum theory of gravity: Syracuse 1949–1962. *The Renaissance of General Relativity in Context*, eds. Blum, A.S., Lalli, R., Renn, J., eds., pp. 221–255 (Springer).

Sanders, R.H. (2014). *Revealing the Heart of the Galaxy: The Milky Way and its Black Hole* (Cambridge University Press).

Sauer, T. (1999). The relativity of discovery: Hilbert's first note on the foundations of physics. *Archive for History of Exact Sciences* 53, 529–575.

Sauer, T. (2005). Einstein Equations and Hilbert Action: What is missing on page 8 of the proofs for Hilbert's First Communication on the Foundations of Physics? *Archive for History of Exact Sciences* 59, 577–590.

Sauer, T. (2014). Marcel Grossmann and his contribution to the general theory of relativity. arXiv:1312.4068.

Scholz, E. (1980). *Geschichte des Mannigfaltigkeitsbegriffs von Riemann bis Poincaré* (Birkhäuser).

Scholz, E. (1999). The concept of manifold, 1850–1950. *History of Topology*, ed. James, I.M., pp. 25–64 (Elsevier).

Scholz, E., ed. (2001). *Hermann Weyl's* Raum - Zeit - Materie *and a General Introduction to His Scientific Work* (Birkhäuser).

Seife, C. (2021). *Hawking Hawking: The Selling of a Scientific Celebrity* (Basic Books).

Senovilla, J.M.M., Garfinkle, D. (2015). The 1965 Penrose singularity theorem. `arXiv:1410.5226`.

Simon, D., ed. (2005). *Albert Einstein: Akademie-Vorträge. Sitzungsberichte der Preußischen Akademie der Wissenschaften 1914–1932* (Wiley–VCH).

Smeenk, C. (2014). Einstein's role in the creation of relativistic cosmology. *The Cambridge Companion to Einstein*, eds. Janssen, M., Lehner, C., pp. 228–269 (Cambridge University Press).

Spivak, M. (1999). *A Comprehensive Introduction to Differential Geometry, Volumes 1-5, Third Edition* (Publish or Perish).

Stachel, J. (1980). Einstein and the rigidly rotating disk. *General Relativity and Gravitation: One Hundred Years After the Birth of Albert Einstein*, Vol. 1, ed. Held, A., pp. 1–15 (Plenum Press).

Stachel, J. (1992). The Cauchy problem in general relativity–The early years. *Studies in the History of General Relativity*, eds. Eisenstaedt, J., Kox, A.J., pp. 407–418 (Birkhäuser).

Stachel, J. (2014). The hole argument and some physical and philosophical implications. *Living Reviews in Relativity* 17, 1-66. `https://link.springer.com/article/10.12942/lrr-2014-1`.

Stuur, S. (2019). *Whose Time Is It? Physicists and Philosophers Debating 'Time' ca. 1900–1930* PhD Thesis, Radboud University Nijmegen. `https://repository.ubn.ru.nl/handle/2066/204263`.

Synge, J.L. (1966). What is Einstein's theory of gravitation? *Perspectives in Geometry and Relativity*, ed. Hoffmann, B., pp. 7–15 (Indiana University Press).

Thorne, K.S. (1994). *Black Holes and Time Warps: Einstein's Outrageous Legacy* (W.W. Norton & Company).

Wightman, A.S. (1976). Hilber's Sixth Problem: Mathematical treatment of the axioms of physics. *Mathematical Developments Arising from Hilbert Problems. Symposia in Pure Mathematics 28*, ed. Browder, F.E., pp. 147–240 (American Mathematical Society).

Wright, A.S. (2013). The origins of Penrose diagrams in physics, art, and the psychology of perception. *Endeavour* 37, 133–139.

Wright, A.S. (2014). The advantages of bringing infinity to a finite place: Penrose diagrams as objects of intuition. *Historical Studies in the Natural Sciences* 44, 99–139.

Books

Abraham, R., Marsden, J.E. (1985). *Foundations of Mechanics*, 2nd ed. (Addison Wesley).

Akivis, M.A., Goldberg, V.V. (1996). *Conformal Differential Geometry and its Generalizations* (Wiley-Interscience).

Anderson, E. (2017). *The Problem of Time* (Springer).

Anderson, J.L. (1967). *Principles of Relativity Physics* (Academic Press).

Armas, J. (2021). *Conversations on Quantum Gravity* (Cambridge University Press).

Ashtekar, A., Berger, B.K., Isenberg, J., MacCallum, M., eds. (2015). *General Relativity and Gravitation: A Centennial Perspective* (Cambridge University Press).

Bär, C., Ginoux, N., Pfäffle, F. (2007). *Wave Equations on Lorentzian Manifolds and Quantization* (European Mathematical Society).

Barbour, J. (1999). *The End of Time* (Oxford University Press).

Barut, A.O., Raçka, R. (1977). *Theory of Group Representations and Applications* (PWN, Warszawa).

Beardon, A.F. (1983). *The Geometry of Discrete Groups* (Springer).

Beem, J.K., Ehrlich, P.E., Easley, K. (1996). *Global Lorentzian Geometry, 2nd Edition* (M. Dekker).

Bergmann, P.G. (1942). *Introduction to the Theory of Relativit* (Prentice-Hall).

Besse, A.L. (1987). *Einstein Manifolds* (Springer).

Blagojević, M., Hehl, F.W. (2013). *Gauge Theories of Gravitation: A Reader with Commentaries* (World Scientific).

Borchers, H.-J., Sen, R.N. (2006). *Mathematical Implications of Einstein–Weyl Causality* (Springer).

Bos, R. (2007). *Groupoids in Geometric Quantization*. PhD Thesis, Radboud University Nijmegen. `https://www.math.ru.nl/~landsman/ProefschriftRogier.pdf`.

Brading, K., Castellani, E. (2003). *Symmetries in Physics: Philosophical Reflections* (Cambridge University Press).

Bravetti, A. (2014). *Geometrothermodynamics: From Ordinary Systems to Black Holes*. PhD Thesis, Sapienza University of Rome.

Brown, H.R. (2005). *Physical Relativity: Space-Time Structure from a Dynamical Perspective* (Oxford University Press).

Callender, C. (2017). *What Makes Time Special?* (Oxford University Press).

Cassirer, E. (1936). Determinismus und Indeterminismus in der modernen Physik: Historische und systematische Studien zum Kausalproblem. *Acta Universitatis Gotoburgensis* XLII, no. 3, pp. 1–256. Reprinted in Cassirer, E. (2004). *Gesammelte Werke, Hamburger Ausgabe, Band 19*. Recki, B. (ed.). Hamburg: Felix Meiner Verlag.

Chadrasekhar, S. (1983). *The Mathematical Theory of Black Holes* (Clarendon Press).

Cheeger, J., Ebin, D.G. (1975). *Comparison Theorems in Riemannian Geometry* (North-Holland).

Choquet-Bruhat, Y. (2009). *General Relativity and the Einstein Equations* (Oxford University Press).

Choquet-Bruhat, Y., DeWitt-Morette, C. (1982). *Analysis, Manifolds and Physics, Revised Edition* (Elsevier).

Christodoulou, D. (2008). *Mathematical Problems of General Relativity I* (European Mathematical Society).

Christodoulou, D. (2009). *The Formation of Black Holes in General Relativity* (European Mathematical Society).

Christodoulou, D., Klainerman, S. (1993). *The Global Nonlinear Stability of the Minkowski Space* (Princeton University Press).

Chruściel, P.T. (2019). *Elements of General Relativity* (Springer).

Chruściel, P.T. (2020). *Geometry of Black Holes* (Oxford University Press).

Clarke, C.J.S. 1993). *The Analysis of Space-Time Singularities* (Cambridge University Press).

Courant, R., Hilbert, D., (1937). *Methoden der mathematischen Physik, Band II: Partielle Differentialgleichungen* (Springer).

Courant, R., Hilbert, D. (1962). *Methods of Mathematical Physics. Volume II: Partial Differential Equations* (Wiley Interscience).

Dainton, B. (2010). *Time and Space. Second Edition* (Acumen Publishing).

Donaldson, S.K., Kronheimer, P.B. (1997). *The Geometry of Four-Manifolds* (Clarendon Press).

Duistermaat, J.J., Kolk, J.C. (2000). *Lie Groups* (Springer).

Earman, J. (1986). *A Primer on Determinism* (Reidel).

Earman, J. (1995). *Bangs, Crunches, Whimpers, and Shrieks: Singularities and Acausalities in Relativistic Spacetimes* (Oxford University Press).

Eisenhart, L.P. (1926). *Riemannian Geometry* (Princeton University Press).

Faraoni, V. (2015). *Cosmological and Black Hole Apparent Horizons* (Springer).

Frankel, T. (2004). *The Geometry of Physics, 2nd Edition* (Cambridge University Press).

Friedman, M. (1983). *Foundations of Space-Time Theories: Relativistic Physics and Philosophy of Science* (Princeton University Press).

Gallot, S., Hulin, D., Lafontaine, J. (1990). *Riemannian Geometry. Second Edition* (Springer).

Gaukroger, S. (2020). *The Failures of Philosophy: A Historical Essay* (Princeton University Press).

Gourgoulhon, E. (2012). *3+1 Formalism in General Relativity - Bases of Numerical Relativity* (Springer).

Giles, R. (1964). *Mathematical Foundations of Thermodynamics* (Pergamon Press).

Grubb, G. (2009). *Distributions and Operators* (Springer).

Guillemin, V., Sternberg, S. (1984). *Symplectic Techniques in Physics* (Cambridge University Press).

Griffiths, J.B., Podolský, J. (2009). *Exact Space-Times in Einstein's General Relativity* (Cambridge University Press).

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

Hawking, S.W., Israel, W., eds. (1979). *General Relativity: An Einstein Centenary Survey* (Cambridge University Press).

Hayward, S.A. (2013). *Black Holes: New Horizons* (World Scientific).

Held, A. (1980). *General Relativity and Gravitation: One Hundred Years After the Birth of Albert Einstein*, Volumes 1 and 2 (Plenum Press).

Helgason, S. (1978). *Differential Geometry, Lie Groups, and Symmetric Spaces* (Academic Press).

Hempel, J. (1976). *3-Manifolds* (Princeton University Press).

Heusler, M. (1996). *Black Hole Uniqueness Theorems* (Cambridge University Press).

Hörmander, L. (1990). *The Analysis of Linear Partial Differential Operators I, Second Edition* (Springer).

Iglesias-Zemmour, P. (2013). *Diffeology* (American Mathematical Society).

Jones, G.A., Singerman, D. (1987). *Complex Functions* (Cambridge University Press).

Joshi, P.S. (1993). *Global Aspects in Gravitation and Cosmology* (Oxford University Press).

Joshi, P.S. (2007). *Gravitational Collapse and Spacetime Singularities* (Cambridge University Press).

Jost, J. (2002). *Riemannian Geometry and Geometric Analysis* (Springer).

Kichenassamy, S. (1996). *Nonlinear Wave Equations* (M. Dekker).

Kirshner, R.P. (2002). *The Extravagant Universe: Exploding Stars, Dark Energy, and the Accelerating Cosmos* (Princeton University Press).

Klainerman, S., Nicolò, F. (2003a). *The Evolution Problem in General Relativity* (Birkhäuser).

Kobayashi, S., Nomizu, K. (1963, 1969). *Foundations of Differential Geometry, Volumes 1, 2* (Wiley).

Krasnov, K. (2020). *Formulations of General Relativity: Gravity, Spinors and Differential Forms* (Cambridge University Press).

Kriele, M. (1999). *Spacetime: Foundations of General Relativity and Differential Geometry* (Springer).

Kijowski, J., Tulczyjew, W.M. (1979). *A Symplectic Framework for Field Theories* (Springer).

Lasky, P. (2010). *Gravitational Collapse in General Relativity: A Unified Treatment* (Lampert).

Lee, D.A. (2019). *Geometric Relativity* (American Mathematical Society).

Lee, J.M. (2012). *Introduction to Smooth Manifolds* (Springer).

Luminet, J.P. (1992). *Black Holes* (Cambridge University Press).

Landsman, K. (1998). *Mathematical Topics Between Classical and Quantum Mechanics* (Springer).

Landsman, K. (2017). *Foundations of Quantum Theory* (Springer).

Laplace, P.S. (1814). *Essai Philosophique sur les Probabilités* (Courcier). (1902). English translation (1902): *A Philosophical Essay on Probabilities* (Wiley).

Malament, D. (2012). *Topics in the Foundations of General Relativity and Newtonian Gravitation Theory* (University of Chicago Press).

Manchak, J.B. (2020). *Global Spacetime Structure* (Cambridge University Press).

Marsden, J.E., Ratiu, T.S. (1999). *Introduction to Mechanics and Symmetry: A Basic Exposition of Classical Mechanical Systems. Second Edition*, (1999).

Martin-Löf, A. (1979). *Statistical Mechanics and the Foundations of Thermodynamics*. Lecture Notes in Physics 101 (Springer).

McTaggart Ellis McTaggart, J. (1921, 1927). *The Nature of Existence, Volumes I, II* (Cambridge University Press).

Misner, C.W., Thorne, K.S., Wheeler, J.A. (1973). *Gravitation* (Freeman).

Moerdijk, I., Mrcun, J. (2003). *Introduction to Foliations and Lie Groupoids* (Cambridge University Press).

Morgan, J.W., Tian, G. (2007). *Ricci Flow and the Poincaré Conjecture* (American Mathematical Society).

Natário, J. (2021). *An Introduction to Mathematical Relativity* (Springer).

Navarro González, J.A., Sancho de Salas, J.B. (2003). *C^∞-Differentiable Spaces* (Springer).

O'Neill, B. (1983). *Semi-Riemannian Geometry* (Academic Press).

O'Neill, B. (1995). *The Geometry of Kerr Black Holes* (AK Peters).

Ortega, J.-P., Ratiu, T.S. (2004). *Momentum Maps and Hamiltonian Reduction* (Birkhäuser).

Penrose, R. (1972). *Techniques of Differential Topology in Relativity* (SIAM).

Penrose, R. (2004). *The Road to Reality: A Complete Guide to the Laws of the universe* (Knopf).

Penrose, R., Rindler, W. (1984). *Spinors and Space-Time. Volume 1: Two-Spinor Calculus and Relativistic Fields* (Cambridge University Press).

Penrose, R., Rindler, W. (1986). *Spinors and Space-Time. Volume 2: Spinor and Twistor Methods in Space-Time Geometry* (Cambridge University Press).

Plebański, J., Krasiński, A. (2006). *An Introduction to General Relativity and Cosmology* (Cambridge University Press).

Poisson, E. (2004). *A Relativist's Toolkit: The Mathematics of Black-Hole Mechanics* (Cambridge University Press).

Rauch, J. (1982). *Hyperbolic Partial Differential Equations and Geometric Optics* (American Mathematical Society).

Rendall, A.D. (2008). *Partial Differential Equations in General Relativity* (Oxford University Press).

Ringström, H. (2009). *The Cauchy Problem in General Relativity* (European Mathematical Society).

Ringström, H. (2013). *On the Topology and Future Stability of the Universe* (European Mathematical Society).

Rovelli, C. (2004). *Quantum Gravity* (Cambridge University Press).

Saint-Gervais, H.P. de (2010). *Uniformization of Riemann Surfaces* (European Mathematical Society).

Scharf, G. (2016). *Gauge Field Theories: Spin One and Spin Two* (Dover).

Shapiro, S. (1991). *Foundations Without Foundationalism: A Case for Second-Order Logic* (Clarendon Press).

Sogge, C.D. (2008). *Lectures on Non-Linear Wave Equations. Second Edition* (International Press).

Souriau, J.-M. (1969). *Structure des Systèmes Dynamiques* (Dunod). Paris. English translation: Souriau, J.-M. (1997). *Structure of Dynamical Systems: a Symplectic View of Physics* (Birkhäuser).

Stephani, H., Kramer, D., MacCallum, M., Hoenselaers, C., Herlt, E. (2003). *Exact Solutions of Einstein's Field Equations. Second Edition* (Cambridge University Press).

Stewart, J. (1991). *Advanced General Relativity* (Cambridge University Press).

Sundermeyer, K. (2014). *Symmetries in Fundamental Physics* (Springer).

Synge, J.L. (1960). *Relativity: The General Theory* (North-Holland).

Taylor, M.E. (1996). *Partial Differential Equations. Volume I: Basic Theory. Volume III: Nonlinear Equations* (Springer).

Thiemann, T. (2007). *Modern Canonical Quantum General Relativity* (Cambridge University Press).

Valiente Kroon, J. (2016). *Conformal Methods in General Relativity* (Cambridge University Press).

Vasy, A. (2015). *Partial Differential Equations* (American Mathematical Society).

Vinberg, E.B., ed. (1993). *Geometry II: Spaces of Constant Curvature* (Springer).

Wald, R.M. 1984). *General Relativity* (University of Chicago Press).

Wald, R.M. (1994). *Quantum Field Theory in Curved Spacetime and Black Hole Thermodynamics* (University of Chicago Press).

Weinberg, S. (1972). *Gravitation and Cosmology: Principles and Applications of the General Theory of Relativity* (Wiley).

Weinberg, S. (2020). *Lectures on Astrophysics* (Cambridge University Press).

Weinstein, T. (1996). *An Introduction to Lorentz Surfaces* (De Gruyter).

Wiltshire, D.L., Visser, M., Scott, S.M., eds. (2009). *The Kerr Spacetime: Rotating Black Holes in General Relativity* (Cambridge University Press).

Wolf, J.A. (2011). *Spaces of Constant Curvature, Sixth Edition* (American Mathematical Society).

Woodhouse, N.M.J. (2007). *General Relativity* (Springer).

Articles and online resources

Adamo, T.M., Newman, E.T., Kozameh C. (2012). Null geodesic congruences, asymptotically-flat spacetimes and their physical interpretation. *Living Reviews in Relativity* 15, 1.

Adams, F.C. (2019). The degree of fine-tuning in our universe–and others. *Physics Reports* 807, 1–111.

Alaee, A., Lesourd, M., Yau, S.-T. (2019). A localized spacetime Penrose inequality and horizon detection with quasi-local mass. `arXiv:1912.01581`.

Alford, F. (2020). The scattering map on Oppenheimer–Snyder space-time. *Annales Henri Poincaré* 21, 2031–2092.

Anderson, A. (2007). On the recovery of geometrodynamics from two different sets of first principles. *Studies in History and Philosophy of Modern Physics* 38, 15–57.

Anderson, I.M. (1981). The principle of minimal gravitational coupling. *Archive for Rational Mechanics and Analysis* 75, 349–372.

Anderson, M.T. (2000a). On stationary vacuum solutions to the Einstein equations. *Annals Henri Poincaré* 1, 977–994.

Anderson, M.T. (2000b). On the structure of solutions to the static vacuum Einstein equations. *Annals Henri Poincaré* 1, 995–1042.

Andersson, L., Mars, M., Simon, W. (2008). Stability of marginally outer trapped surfaces and existence of marginally outer trapped tubes. *Advances in Theoretical and Mathematical Physics* 12, 853–888.

Andersson, L., Metzger, J. (2009). The area of horizons and the trapped region. *Communications in Mathematical Physics* 290, 941–972.

Andrews, B. (undated). *Lectures on Differential Geometry*. `http://maths-people.anu.edu.au/~andrews/DG/`.

Aretakis, S. (2013). *Lecture Notes on General Relativity*. `https://web.math.princeton.edu/~aretakis/columbiaGR.pdf`.

Aretakis, S., Rodnianski, I. (2015). The Cauchy problem in General Relativity. *General Relativity and Gravitation: A Centennial Perspective*, eds. Ashtekar, A. *et al.*, pp. 452–479 (Cambridge University Press).

Ashtekar, A. (1980). Asymptotic structure of the gravitational field at spatial infinity. *General Relativity and Gravitation: One Hundred Years After the Birth of Albert Einstein*, Vol. 2, ed. Held, A., pp. 37–70 (Plenum Press).

Ashtekar, A. (2015). Geometry and physics at null infinity. `arXiv:1409.1800`.

Ashtekar, A., Bonga, B., Kesavan, A. (2015). Asymptotics with a positive cosmological constant: I. Basic framework. *Classical and Quantum Gravity* 32, 025004.

Ashtekar, A., Galloway, G.J. (2005). Some uniqueness results for dynamical horizons. *Advances in Theoretical and Mathematical Physics* 9, 1–30.

Ashtekar, A., Krishnan, B. (2004). Isolated and dynamical horizons and their applications. *Living Reviews in Relativity* 7:10.

Ashtekar, A., Magnon, A. (1984). Asymptotically anti-de Sitter space-times. *Classical and Quantum Gravity* 1, L39–L44.

Athanasiou, N., Lesourd, M. (2020). Construction of Cauchy data for the dynamical formation of apparent horizons and the Penrose Inequality. `arXiv:2009.03704`.

Bär, C. (2007/08). *Geometrische Analysis*. `https://www.math.uni-potsdam.de/fileadmin/user_upload/Prof-Geometrie/Dokumente/Lehre/Lehrmaterialien/skript-GeomAna.pdf`.

Bardeen, J.M., Carter, B., Hawking, S.W. (1973). The four laws of black hole mechanics. *Communications in Mathematical Physics* 31, 161–170.

Bartnik, R., Isenberg, J. (2004). The constraint equations *The Einstein Equations and the Large Scale Scale Behaviour of Gravitational Fields*, eds. Friedrich, H., Chruściel, P.T., pp. 1–38 (Springer).

Beig, R. (1980). The static gravitational field near spatial infinity I. *General Relativity and Gravitation* 12, 439–451.

Beig, R., Simon, W. (1980). The stationary gravitational field near spatial infinity. *General Relativity and Gravitation* 12, 1003–1013.

Bekenstein, J. D. (1972). Black Holes and the Second Law. *Lettere al Nuovo Cimento* 4, 737–740.

Bekenstein, J.D. (1973). Black holes and entropy. *Physical Review D* 7, 2333–2346.

Bekenstein, J. D. (1974). Generalized second law of thermodynamics in black-hole physics. *Physical Review D* 9, 3292–3300.

Belot, G. (2013). Symmetry and equivalence. *The Oxford Handbook of Philosophy of Physics*, ed. Batterman, R., pp. 318–339 (Oxford University Press).

Ben-Dov, I. (2004). The Penrose inequality and apparent horizons. *Physical Review D* 70:124031.

Bernal, A. N., Sánchez, M. (2003). On smooth Cauchy hypersurfaces and Geroch's splitting theorem. *Communications in Mathematical Physics* 243, 461–470.

Bernal, A. N., Sánchez, M. Smoothness of time-functions and the metric splitting of globally hyperbolic spacetimes. *Communications in Mathematical Physics* 257, 43–50.

Bernal, A. N., Sánchez, M. (2006a). Further results on the smoothability of Cauchy hypersurfaces and Cauchy time functions, *Letters in Mathematical Physics* 77, 183–197.

Bernal, A. N., Sánchez, M. (2006b). Globally hyperbolic spacetimes can be defined as "causal" instead of "strongly causal". *Classical and Quantum Gravity* 24, 745–749.

Bieri, L. (2018). Black hole formation and stability: A mathematical investigation. *Bulletin of the American Mathematical Society (N.S.)* 55, 1–30.

Blohmann, C., Barbosa Fernandes, M.C., Weinstein, A. (2013). Groupoid symmetry and constraints in general relativity. *Communications in Contemporary Mathematics* 15:1250061.

Blohmann, C., Weinstein, A. (2018). Hamiltonian Lie algebroids. `arXiv:1811.11109`. *Memoirs of the American Mathematical Society*, to appear.

Bojowald, M., Büyükąm, U., Brahma, S., D'Ambrosio, F. (2016). Hypersurface-deformation algebroids and effective spacetime models. *Physical Review D* 94:104032

Booth, I. (2005). Black hole boundaries. *Canadian Journal of Physics* 83, 1073–1099.

Born, M. (1926). Zur Quantenmechanik der Stoßvorgänge. *Zeitschrift für Physik* 37, 863–867.

Bray, H.L. (2001). Proof of the Riemannian Penrose inequality using the positive mass theorem. *Journal of Differential Geometry* 59, 177–267.

Bray, H.L. (2002). Black holes, geometric flows, and the Penrose inequality in general relativity. *Notices of the AMS* 49, 1372–1381.

Bray, H.L., Chruściel, P.T. (2004). The Penrose inequality. *The Einstein Equations and the Large Scale Behavior of Gravitational Fields*, eds. Friedrich, H., Chruściel, P.T., pp. 39–70 (Springer).

Browdy, S.F., Galloway, G.J. (1995). Topological censorship and the topology of black holes. *Journal of Mathematical Physics* 36, 4952–4961.

Budic, R., Isenberg, J., Lindblom, L., Yasskin, P.B. (1978). On the determination of Cauchy surfaces from intrinsic properties. *Communications in Mathematical Physics* 61, 87–95.

Bunting, G.L., Masood-ul-Alam, A.K.M. (1987). Nonexistence of multiple black holes in asymptotically Euclidean static vacuum space-time. *General Relativity and Gravitation* 19, 147–154.

Burtscher, A.Y. (2020). Initial data and black holes for matter models. *Hyperbolic Problems: Theory, Numerics, Applications. AIMS Series in Applied Mathematics* 10, 336–345.

Burtscher, A.Y., LeFloch, P.G. (2014). The formation of trapped surfaces in spherically-symmetric Einstein–Euler spacetimes with bounded variation. *Journal de Mathématiques Pures et Appliquées* 102, 1164–1217.

Butterfield, J. (1984). Seeing the present. *Mind* 93, 161–176.

Butterfield, J. (1987). Substantivalism and determinism. *International Studies in the Philosophy of Science* 2, 10–32.

Butterfield, J. (1989). The hole truth. *British Journal for the Philosophy of Science* 40, 1989, 1–28.

Candela, A.M., Flores, J.L., Sánchez, M. (2010). Global hyperbolicity and Palais–Smale condition for action functionals in stationary spacetimes. *Advances in Mathematics* 218, 515–536.

Candela, A.M., Sánchez, M. (2010). Geodesics in semi-Riemannian manifolds: Geometric properties and variational tools. *Recent Developments in Pseudo-Riemannian Geometry*, ed. Alekseevskii, D.V., pp. 359–418 (European Mathematical Society).

Carlip, S (2014). Black hole thermodynamics. *International Journal of Modern Physics* D23, 1430023–736.

Carlotto, A, (2021). The general relativistic constraint equations. *Living Reviews in Relativity* 24:2.

Carter, B. (1971a). Causal structure in space-time. *General Relativity and Gravitation* 1, 349–391.

Carter, B. (1971b). Axisymmetric black hole has only two degrees of freedom. *Physical Review Letters* 26, 331–333.

Carter, B. (1973). Black hole equilibrium states. Part I: Analytic and geometric properties of the Kerr solution. *Black Holes–Les astres occlus*, eds. De Witt, B., DeWitt-Morette, C., pp. 61–124 (Gordon and Breach). Reprinted in *General Relativity and Gravitation* 41, 2873–2938 (2009).

Carter, B. (1979). The general theory of the mechanical electromagnetic, and thermodynamic properties of black holes. *General Relativity: An Einstein Centenary Survey*, eds. Hawking, S.W., Israel, W., pp. 274–369 (Cambridge University Press).

Carter, B. (1986). Mathematical foundations of the theory of relativistic stellar and black hole configurations. *Gravitation in Astrophysics: Cargèse 1986*, eds. Carter, B., Hartle, J.B., pp. 63–122 (Plenum Press).

Castelvecchi, D. (2020). Astronomers detect mindboggling black-hole collision. *Nature* 585, 171–172.

Caulton, A. (2015). The role of symmetry in the interpretation of physical theories. *Studies in History and Philosophy of Modern Physics* 52, 153–162.

Cederbaum, C. (2019). *Static black hole uniqueness theorems. Lectures 1–4. ICTP School of Geometry and Gravity*. `https://www.youtube.com/watch?v=hf4qIiGVwLk` etc.

Cederbaum, C., Sakovich, A. (2018). On center of mass and foliations by constant spacetime mean curvature surfaces for isolated systems in General Relativity. `arXiv:1901.00028`.

Chernov, V., Nemirovski, S. (2013). Cosmic censorship of smooth structures. *Communications in Mathematical Physics* 320, 469–473.

Chesler, P.M., Narayan, R., Curiel, E. (2020). Singularities in Reissner–Nordström black holes. *Classical and Quantum Gravity* 37, 025009.

Choquet-Bruhat, Y., Chruściel, P.T., Martín-García, J.M. (2011). The Cauchy problem on a characteristic cone for the Einstein equations in arbitrary dimensions. *Annales Henri Poincaré*, 12, 419–482.

Choquet-Bruhat, Y., York, J.W. (1980). The Cauchy problem. *General Relativity and Gravitation: One Hundred Years After the Birth of Albert Einstein*, Vol. 1. Held, A., ed., pp. 99–172 (Plenum Press).

Christodoulou, D. (1970). Reversible and irreversible transformations in black-hole physics. *Physical Review Letters* 25, 1596–1597.

Christodoulou, D., Ruffini, R. (1971). Reversible transformations of a charged black hole. *Physical Review D* 3552–3555.

Christodoulou, D. (1991). The formation of black holes and singularities in spherically symmetric gravitational collapse. *Communications in Pure and Applied Mathematics* 44, 339–373.

Christodoulou, D. (1999a). On the global initial value problem and the issue of singularities. *Classical and Quantum Gravity* 16, A23–A35.

Christodoulou, D. (1999b). The instability of naked singularities in the gravitational collapse of a scalar field. *Annals of Mathematics* 149, 183–217.

Chruściel, P.T. (1992). On uniqueness in the large of solutions of Einstein's equations ("Strong Cosmic Censorship"). *Mathematical Aspects of Classical Field Theory*, eds. Gotay, M.J., Marsden, J.E., Moncrief, V., pp. 235–274. *Contemporary Mathematics* 132 (American Mathematical Society).

Chruściel, P.T. (1996). Uniqueness of stationary, electro-vacuum black holes revisited. `arXiv:gr-qc/9610010`.

Chruściel, P.T. (1997). On rigidity of analytic black holes. *Communications in Mathematical Physics* 189, 1–7.

Chruściel, P.T. (2002). Black holes. *Springer Lecture Notes in Physics* 604, 61–102.

Chruściel, P.T. (2010). *An introduction to the Cauchy problem for the Einstein equations*. `https://homepage.univie.ac.at/piotr.chrusciel/teaching/Cauchy/Roscoff.pdf`.

Chruściel, P.T. (2011). Elements of causality theory. `arXiv:1110.6706.pdf`.

Chruściel, P.T. (2014). On maximally globally hyperbolic vacuum space-times. *Journal of Fixed Point Theory and Applications* 14, 325–353.

Chruściel, P.T., Delay, E. (2002). Existence of non-trivial, vacuum, asymptotically simple spacetimes. *Classical and Quantum Gravity* 19, L71–L79.

Chruściel, P.T., Delay, E., Galloway, G.J., Howard, R. (2001). Regularity of horizons and the area theorem. *Annales Henri Poincaré* 2, 109–178.

Chruściel, P.T., Galloway, G.J. (2014). Outer trapped surfaces are dense near MOTSs. *Classical and Quantum Gravity* 31:045013.

Chruściel, P.T., Galloway, G.J. (2019). Roads to topological censorship. `arXiv:1906.02151`.

Chruściel, P.T., Galloway, G.J., Pollack, D. (2010). Mathematical general relativity: A sampler. *Bulletin of the American Mathematical Society* 47, 567–638.

Chruściel, P.T., Grant, J.D.E., and Minguzzi, E. (2016). On differentiability of volume time functions. *Annales Henri Poincaré* 17, 2801–2824.

Chruściel, P.T., Lopes Costa, J. (2008). On uniqueness of stationary black holes. *Astérisque* 321, 195–265.

Chruściel, P.T., Lopes Costa, J., Heusler, M. (2012). Stationary black holes: Uniqueness and beyond. *Living Reviews in Relativity* 15:7. `https://link.springer.com/article/10.12942/lrr-2012-7`.

Chruściel, P.T., Nguyen, L. (2010). A uniqueness theorem for degenerate Kerr–Newman black holes. *Annales Henri Poincaré* 11, 585–609.

Chruściel, P.T., Paetz, T.T. (2012). The many ways of the characteristic Cauchy problem. *Classical and Quantum Gravity* 29:145006.

Chruściel, P.T., Paetz, T.T. (2015). Characteristic initial data and smoothness of Scri. I. Framework and results. *Annales Henri Poincaré* 16, 2131–2162.

Chruściel, P.T., Wald, R.M. (1994a). On the topology of stationary black holes. *Classical and Quantum Gravity* 11, L147–L152.

Chruściel, P.T., Wald, R.M. (1994b). Maximal hypersurfaces in stationary asymptotically flat spacetimes. *Communications in Mathematical Physics* 163, 561–604.

Clarke, C.J.S., Joshi, P.S. (1988). On reflecting spacetimes. *Classical and Quantum Gravity* 5, 19–25.

Claudel, C.-M. (2000). Black holes and closed trapped surfaces: A revision of a classic theorem. `arXiv:gr-qc/0005031`.

Coley, A.A. (2019). Mathematical general relativity. *General Relativity and Gravitation* 51:78.

Compère, G. (2006). An introduction to the mechanics of black holes. `arXiv:gr-qc/0611129`.

Cortier, J., Minerbe, V. (2016). On complete stationary vacuum initial data. *Journal of Geometry and Physics* 99, 20–27.

Cotton, É. (1899). Sur les variétés a trois dimensions. *Annales de la Faculté des sciences de Toulouse (2^e série)* 1, 385–438.

Corvino, J. (2007). On the existence and stability of the Penrose compactification. *Annales de l'Institut Henri Poincaré* 8, 597–620.

Corvino, J., Pollack, D. (2011). Scalar curvature and the Einstein constraint equations. *Surveys in Geometric Analysis and Relativity*, eds. Bray, H.L., Minicozzi, W.P., pp. 145–188 (International Press).

Crisford, T., Santos, J.E. (2017). Violating the weak cosmic censorship conjecture in four-dimensional Anti–de Sitter space. *Physical Review Letters* 118:181101.

Curiel, E. (2014a). A primer on energy conditions. `arXiv:1405.0403`.

Curiel, E. (2014b). Are classical black holes hot or cold? `http://philsci-archive.pitt.edu/11136/1/class-bhs-hot.pdf`.

Curiel, E. (2019a). Singularities and Black Holes. *The Stanford Encyclopedia of Philosophy (Spring 2019 Edition)*, ed. Zalta, E.N. `https://plato.stanford.edu/archives/spr2019/entries/spacetime-singularities/`.

Curiel, E. (2019b). The many definitions of a black hole. *Nature Astronomy* 3, 27–34.

Dafermos, M. (2003). Stability and instability of the Cauchy horizon for the spherically symmetric Einstein–Maxwell-scalar field equations. *Annals of Mathematics* 158, 875–928.

Dafermos, M. (2005). The interior of charged black holes and the problem of uniqueness in general relativity. *Communications on Pure and Applied Mathematics* 58, 445–504.

Dafermos, M. (2009). The evolution problem in general relativity. *Current Developments in Mathematics*, eds. Jerison, D. *et al.*, pp. 1–66 (International Press).

Dafermos, M. (2012). The formation of black holes in General Relativity [after D. Christodoulou]. *Séminaire Bourbaki* 64, no. 1051. https://www.dpmms.cam.ac.uk/~md384/expose-chr.pdf.

Dafermos, M. (2014a). Black holes without spacelike singularities. *Communications in Mathematical Physics* 332, 729–757.

Dafermos, M. (2014b). The mathematical analysis of black holes in general relativity. *Proceedings of the ICM, 2014*. https://www.dpmms.cam.ac.uk/~md384/ICMarticleMihalis.pdf.

Dafermos, M. (2017). The cosmic censorship conjectures in classical general relativity. https://www.youtube.com/watch?v=ZBYAbejIvB4.

Dafermos, M. (2019). The cosmic censorship conjectures in general relativity (ICTP School on Geometry and Gravity). Lecture 1: https://www.youtube.com/watch?v=Lg1Cetf7V9I.
Lecture 2: https://www.youtube.com/watch?v=SoRhBSt_mN0.

Dafermos, M., Holzegel, G., Rodnianski, I. (2019a). The linear stability of the Schwarzschild solution to gravitational perturbations. *Acta Mathematica* 222, 1–214.

Dafermos, M., Holzegel, G., Rodnianski, I. (2019b). Boundedness and decay for the Teukolsky equation on Kerr spacetimes I: The Case $|a| \ll M$. *Annals of PDE* 5:2.

Dafermos, M., Holzegel, G., Rodnianski, Taylor, M. (2021). The non-linear stability of the Schwarzschild family of black holes. arXiv:2104.08222.

Dafermos, M., Rodnianski, I. (2008). Lectures on black holes and linear waves. arXiv:0811.0354.

D'Ambra, G., Gromov, M. (1991). Lectures on transformation groups: Geometry and dynamics. *Surveys in Differential Geometry* 1, 19–111.

De Haro, S. (2021). Noether's theorems and energy in general relativity. arXiv:2103.17160.

Deser, S., Franklin, J. (2005). Schwarzschild and Birkhoff a la Weyl. *American Journal of Physics* 73, 261–264.

Dewar, N. (2019). Sophistication about symmetries. *British Journal for the Philosophy of Science* 70, 485–521.

Dewar, N. (2020). General-Relativistic covariance. *Foundations of Physics* 50, 294–318.

Dias, O.J.C., Reall, H.S., Santos, J.E. (2018a). Strong cosmic censorship: Taking the rough with the smooth. *Journal of High Energy Physics* 2018:1.

Dias, O.J.C., Eperon, F.C., Reall, H.S., Santos, J.E. (2018b). Strong cosmic censorship in de Sitter space. *Physical Review D* 97:104060.

Doboszewski, J. (2017). Non-uniquely extendible maximal globally hyperbolic spacetimes in classical general relativity: A philosophical survey. *European Studies in Philosophy of Science* 6, 193– 212.

Doboszewski, J. (2019). Relativistic spacetimes and definitions of determinism. *European Journal for Philosophy of Science* 9:24.

Doboszewski, J. (2020). Epistemic holes and determinism in classical general relativity. *British Journal for the Philosophy of Science* 71, 1093–1111.

Dougherty, J., Callender, G. (2016). Black hole thermodynamics: More than an analogy? http://philsci-archive.pitt.edu/13195/1/bht.pdf.

Earman, J. (2002). Thoroughly modern McTaggart: Or, what McTaggart would have said if he had read the general theory of relativity. *Philosophers Imprint* 2, 1–28.

Earman, J. (1996). Tolerance for spacetime singularities. *Foundations of Physics* 26, 623–640.

Earman, J. (2006a). Two challenges to the requirement of substantive general covariance. *Synthese* 148, 443–468.

Earman, J. (2006b). The implications of general covariance for the ontology and ideology of spacetime. *The Ontology of Space-Time*, ed. Dieks, D., pp. 3–24 (Elsevier).

Earman, J. (2007). Aspects of determinism in modern physics. *Handbook of the Philosophy of Science. Vol. 2: Philosophy of Physics, Part B*, eds. Butterfield, J., Earman, J., pp. 1369–1434 (North-Holland).

Earman, J., Norton, J.D. (1987). What price substantivalism? The hole story. *British Journal for the Philosophy of Science* 9, 251–278.

Eichmair, M., Galloway, G.J., Pollack, D. (2013). Topological censorship from the initial data point of view *Journal of Differential Geometry* 95, 389–405.

Event Horizon Telescope Collaboration (2019a). First M87 event horizon telescope results. I. The shadow of the supermassive black hole. *The Astrophysical Journal Letters* 875:L1, 1–17.

Event Horizon Telescope Collaboration (2019b). First M87 event horizon telescope results. V. Physical origin of the asymmetric ring. *The Astrophysical Journal Letters* 875:L5, 1–31.

Falcke, H., Melia, F., Agol, E. (1999). Viewing the shadow of the black hole at the galactic center. *The Astrophysical Journal Letters* 528, L13–L16.

Fathi, A., Siconolfi, A. (2012). On smooth time functions. *Mathematical Proceedings of the Cambridge Philosophical Society* 152, 303–339.

Fewster, C.J., Galloway, G.J. (2011). Singularity theorems from weakened energy conditions. *Classical and Quantum Gravity* 28:125009.

Fewster, C.J., Kontou, E. (2020). A new derivation of singularity theorems with weakened energy hypotheses. *Classical and Quantum Gravity* 37:065010.

Fischer, A.E., Marsden, J.E. (1979). The initial value problem and the dynamical formulation of general relativity. *General Relativity: An Einstein Centenary Survey*, eds. Hawking, S.W., Israel, W., pp. 138–211 (Cambridge University Press).

Fischer, A.E., Moncrief, V. (1996). A method of reduction of Einstein's equations of evolution and a natural symplectic structure on the space of gravitational degrees of freedom. *General Relativity and Gravitation* 28, 207–19.

Fischer, A.E., Moncrief, V. (1997). Hamiltonian reduction of Einstein's equations of general relativity. *Nuclear Physics B Proceedings Supplements* 57, 142–161.

Forger, M., Romero, S.V. (2005). Covariant Poisson brackets in geometric field theory. *Communications in Mathematical Physics* 256, 375–410.

Frauendiener, J. (2000). Conformal infinity. *Living Reviews in Relativity* 3:4.

Freivogel, B., Kontou, E.A., Krommydas, D. (2020). The return of the singularities: Applications of the smeared null energy condition. `arXiv:2012.11569`.

Friedman, J.L., Schleich, K., Witt, D.M. (1993). Topological censorship. *Physical Review Letters* 71, 1486–1489. Erratum *ibid.* 75, 1872 (1995).

Friedrich, H. (1979). *Eine Untersuchung der Einsteinschen Vakuumfeldgleichungen in der Umgebung regulärer und singulärer Nullhyperflächen*, PhD Thesis, University of Hamburg.

Friedrich, H. (2004). Smoothness at null infinity and the structure of initial data. *The Einstein Equations and the Large Scale Behavior of Gravitational Fields*, eds. Friedrich, H., Chruściel, P.T., pp. 121–203 (Springer).

Friedrich, H. (2018). Peeling or not peeling–Is that the question? *Classical and Quantum Gravity* 35:083001.

Friedrich, H., Rácz, I., Wald, R.M. (1999). On the rigidity theorem for space-times with a stationary event horizon or a compact Cauchy horizon. *Communications in Mathematical Physics* 204, 691–707.

Friedrich, H., Rendall, A.D. (2000). *The Cauchy Problem for the Einstein equations.* `arXiv:gr-qc/0002074`.

Friedrich, H., Stewart, J.M. (1983). Characteristic initial data and wave front singularities in general relativity. *Proceedings of the Royal Society of London A* 385, 345–371.

Gajic, D., Luk, J. (2019). The interior of dynamical extremal black holes in spherical symmetry. *Pure and Applied Analysis* 1, 263–326.

Galloway, G.J. (1985). Some results on Cauchy surface criteria in Lorentzian geometry. *Illinois Journal of Mathematics* 29, 1–10.

Galloway, G.J. (1995). On the topology of the domain of outer communication. *Classical and Quantum Gravity* 12, L99–L101.

Galloway, G.J. (2014). Notes on Lorentzian causality. `https://www.math.miami.edu/~galloway/vienna-course-notes.pdf`.

Galloway, G.J. (2017). Topology and general relativity. `https://www.math.miami.edu/~galloway/ESI2017.pdf`. See also `https://www.youtube.com/watch?v=3z2oTCoBHVo`.

Galloway, G.J., Ling, E. (2017). Some remarks on the C^0-(in)extendibility of spacetimes. *Annales Henri Poincaré* 18, 3427–3447.

Galloway, G.J., Ling, E., Sbierski, J. (2018). Timelike completeness as an obstruction to C^0-extensions. *Communications in Mathematical Physics* 359, 937–949.

Galloway, G.J., Miao, P., Schoen, R. (2015). Initial data and the Einstein constraint equations. Ashtekar, A. *et al.*, eds., pp. 412–448 (Cambridge University Press).

Gannon, D. (1975). Singularities in nonsimply connected space-times. *Journal of Mathematical Physics* 16, 2364–2367.

Gao, S., Wald, R.M. (2001). "Physical process version" of the first law and the generalized second law for charged and rotating black holes. *Physical Review D* 64, 084020.

Garcia, A., Hehl, F.W., Heinicke, C., Macias, A. (2004). The Cotton tensor in Riemannian spacetimes. *Classcal & Quantum Gravity* 21, 1099–1118.

Geroch, R. (1977). Asymptotic structure of space-time. *Asymptotic Structure of Space-Time*, eds. Esposito, F.P., Witten, L., pp. 1–105 (Plenum).

Geroch, R., Horowitz, G. (1978). Asymptotically simple does not imply asymptotically Minkowskian. *Physical Review Letters* 40, 203–206.

Geroch, R., Horowitz, G. (1979). Global structure of spacetimes. *General Relativity: An Einstein Centenary Survey*, eds. S.W. Hawking and W. Israel, pp. 212–293 (Cambridge University Press).

Geroch, R., Jang, P.S. (1975). Motion of a body in general relativity, *Journal of Mathematical Physics* 16, 65–67.

Geroch, R., Kronheimer, E.H., Penrose, R. (1972). Ideal points in space-time. *Proceedings of the Royal Society (London)* A327, 545–567.

Geroch, R., Traschen, J. (1987). Strings and other distributional sources in general relativity. *Physical Review D* 36, 1017–1031.

Geroch, R., Weatherall, J.O. (2018). The motion of small bodies in space-time. *Communications in Mathematical Physics* 364, 607–634.

Giorgi, E., Klainerman, S., Szeftel, J. (2020). A general formalism for the stability of Kerr. `arXiv:2002.02740`.

Głowacki, J. (2019). *Groupoid Symmetry and Constraint Bracket of General Relativity Revisited*. M.Sc Thesis, Radboud University Nijmegen. `https://www.math.ru.nl/~landsman/Jan2019.pdf`.

Głowacki, J. (2021). Inevitability of the Poisson bracket structure of the relativistic constraints. *Foundations of Physics*, under review.

Gomes, H., Butterfield, J. (2020). Geometrodynamics as functionalism about time. `arXiv:2010.16199`.

Gomes, H., Shyam, V. (2016). Extending the rigidity of general relativity. *Journal of Mathematical Physics* 57, 112503.

Gotay, M.J., Isenberg, J., Marsden, J.E., Montgomery, R., with the collaboration of Śniatycki. J., Yasskin, P.B. (1998–2004). Momentum maps and classical relativistic fields. Part I: Covariant field theory. `arXiv:physics/9801019`. Part II: Canonical analysis of field theories. `arXiv:math-ph/041103`.

Graf, M., Grant, J.D.E., Kunzinger, M., Steinbauer, R. (2018). The Hawking–Penrose Singularity Theorem for $C^{1,1}$ Lorentzian Metrics. *Communications in Mathematical Physics* 360, 1009–1042.

Graves, J.C., Brill, D.R. (1960). Oscillatory character of Reissner-Nordström metric for an ideal charged wormhole. *Physical Review* 120, 1507–1513.

Grieser, D. (2000). Basics of the *b*-calculus. `arXiv:math/0010314`.

Gryb, S., Thébault, K. (2016). Time remains. *British Journal for the Philosophy of Science* 67, 663–705.

Häfner, D., Hintz, P., Vasy, A. (2019). Linear stability of slowly rotating Kerr black holes. `arXiv:1906.00860`.

Hartman, P. (1983). Remarks on geodesics. *Proceedings of the American Mathematical Society* 89, 467–472.

Hartman, P., Wintner, A. (1951). On the problems of geodesics in the small. *American Journal of Mathematics* 73, 132–148.

Hawking, S.W. (1972). Black holes in general relativity. *Communications in Mathematical Physics* 25, 152–166.

Hawking, S.W. (1974). Black hole explosions? *Nature* 248 (5443), 30–31.

Healey, R. (2002). Can physics coherently deny the reality of time? *Royal Institute of Philosophy Supplement* 50, 293–316.

Healey, R. (2004). Change without change, and how to observe it in general relativity. *Synthese* 141, 381–415.

Heckman, G. (2017). *Introduction to Riemannian Geometry*. `https://www.math.ru.nl/~heckman/DiffGeom.pdf`.

Heinicke, C., Hehl, F.W. (2015). Schwarzschild and Kerr solutions of Einstein's field equation–an introduction. `arXiv:1503.02172`.

Henry, R.C. (2000). Kretschmann scalar for a Kerr–Newman black hole. *The Astrophysical Journal* 535, 350–353.

Hertog, T., Horowitz, G.T., Maeda, K. (2004). Generic Cosmic-Censorship violation in anti–de Sitter Space. *Physical Review Letters* 92:131101.

Hintz, P., Vasy, A. (2018). The global non-linear stability of the Kerr–de Sitter family of black holes. *Acta Mathematica* 220, 1–206.

Hiscock, W.A. (1981). Evolution of the interior of a charged black hole. *Physics Letters A* 83, 110–112.

Hojman, S.A., Kuchar, K., Teitelboim, C. (1976). Geometrodynamics regained. *Annals of Physics* 96, 88–135.

Holst, M., Maxwell, D., Mazzeo, R. (2017). Conformal fields and the structure of the space of solutions of the Einstein constraint equations. `arXiv:1711.01042`.

Hounnonkpe, R. A., Minguzzi, E. (2019). Globally hyperbolic spacetimes can be defined without the 'causal' condition. `arXiv:1908.11701`.

Horowitz, G. (1979). Finding a statement of cosmic censorship. *General Relativity and Gravitation* 10, 1057–1061.

Huisken, G., Ilmanen, T. (1997). The Riemannian Penrose inequality. *International Mathematical Research Notices* 20, 1-45–1058.

Huisken, G., Ilmanen, T. (2001). The inverse mean curvature flow and the Riemannian Penrose inequality. *Journal of Differential Geometry* 59, 353–437.

Isenberg, J. (2014). The initial value problem in general relativity *Springer Handbook of Spacetime*, eds. Ashtekar, A., Petkov, V., pp. 303–321 (Springer).

Isenberg, J., Nester, J. (1980). Canonical gravity. *General Relativity and Gravitation: One Hundred Years After the Birth of Albert Einstein*. Vol. 1, ed. Held, A., pp. 23–98 (Plenum Press).

Isham, C.J. (1992). Canonical quantum gravity and the problem of time. *Integrable Systems, Quantum Groups, and Quantum Field Theory*, eds. Ibort, L.A., Rodriquez, M.A., pp. 157–287 (Kluwer).

Isham, C.J., Kuchar, K.V. (1985a). Representations of spacetime diffeomorphisms. I. Canonical parametrized field theories. *Annals of Physics* 164, 288–315.

Isham, C.J., Kuchar, K.V. (1985b). Representations of spacetime diffeomorphisms. II. Canonical geometrodynamics. *Annals of Physics* 164, 316-333.

Israel, W. (1986). Third law of black-hole dynamics: A formulation and proof. *Physical Review Letters* 57, 397–399.

Iyer, V., Wald, R.M. (1994). Some properties of the Noether charge and a proposal for dynamical black hole entropy. *Physical Review D* 50, 846–864.

Jacobson, T. (1996). Introductory lectures on black hole thermodynamics. `https://fac.ksu.edu.sa/sites/default/files/t_jackbson_lecture_notes_on_black_hole_thermodynamics.pdf`.

Jacobson, T., Kang, G., Myers, R.C. (1994). On black hole entropy. *Physical Review D* 49, 6587–6598.

Jacobson, T., Venkataramani, S. (1995). Topology of event horizons and topological censorship. *Classical and Quantum Gravity* 12, 1055–1061.

Jaramillo, J.L., Gourgoulhon, E. (2009). Mass and angular momentum in general relativity. *Mass and Motion in General Relativity*, eds. Blanchet, L., Spallicci, A., Whiting, B., pp. 87–124 (Springer).

Joshi, P.S. (2014). *Spacetime singularities*, `arXiv:1311.0449`.

Kaiser, D. (2012). A tale of two textbooks: Experiments in genre. *Isis* 103, 126–138.

Kazdan, J.L. (1981). Another proof of Bianchi's identity in Riemannian geometry, *Proceedings of the American Mathematical Society* 81, 341–342.

Kennefick, D., Ó Murchadha, N. (1995). Weakly decaying asymptotically flat static and stationary solutions to the Einstein equations. *Classical and Quantum Gravity* 12, 149–158.

Klainerman, S. (2014). Are black holes real? `https://www.youtube.com/watch?v=zj1QkhvHVGU`.

Klainerman, S., Luk, J., Rodnianski, I. (2014). A fully anisotropic mechanism for formation of trapped surfaces. *Inventiones Mathematicae* 198, 1–26.

Klainerman, S., Nicolò, F. (2003). Peeling properties of asymptotically flat solutions to the Einstein vacuum equations. *Classical and Quantum Gravity* 20, 3215–3258.

Klainerman, S., Rodnianski, I. (2012). On the formation of trapped surfaces. *Acta Mathematica* 208, 211–213.

Klainerman, S., Rodnianski, I., Szeftel, J. (2015). The bounded L^2 curvature conjecture. *Inventiones Mathematicae* 202, 91–216.

Klainerman, S., Szeftel, J. (2017). Global nonlinear stability of Schwarzschild spacetime under polarized perturbations. `arXiv:1711.07597`.

Klainerman, S., Szeftel, J. (2021). Kerr stability for small angular momentum. `arXiv:2104.11857`.

Kostant, B. (1970). Quantization and unitary representations. *Lecture Notes in Mathematics* 170, 87–208.

Kragh, H. (2016). Ludvig Lorenz, electromagnetism, and the theory of telephone currents. `arXiv:1606.00205`.

Królak, A. (1986). Towards the proof of the cosmic censorship hypothesis. *Classical and Quantum Gravity* 3, 267–280.

Królak, A. (1999). Nature of singularities in gravitational collapse. *Progress of Theoretical Physics Supplement* 136, 45–56.

Królak, A. (2004). Cosmic censorship hypothesis. *Contemporary Mathematics* 359, 51–64.

Kuchar, K. (1976). Geometry of hyperspace. I. *Journal of Mathematical Physics* 17, 777–791.

Kuchar, K. (1992). Time and interpretations of quantum gravity. *Proceedings of the 4th Canadian Conference on General Relativity and Relativistic Astrophysics*, eds. Kunstatter, G., Vincent, D., Willams, J., pp. 1–104 (World Scientific).

Künzle, H.P. (1971). On the spherical symmetry of a static perfect fluid. *Communications in Mathematical Physics* 20, 85–100.

Kupeli, D.N. (1987). On null submanifolds in spacetimes. *Geometriae Dedicata* 23, 33–51.

Landsman, K. (2016). The Fine-Tuning Argument. *The Challenge of Chance*, eds. Landsman, K., van Wolde, E., pp. 111–130 (Springer).

Landsman, K. (2020). Randomness? What randomness? *Foundations of Physics* 50, 61–104 (2020).

Landsman, K. (2021). Indeterminism and undecidability. *Undecidability, Uncomputability, and Unpredictability*, eds. Aguirre, A., Merali, Z., Sloan, D., pp. 17–45 (Springer). `arXiv:2003.03554`.

Landsman, K., Wiedemann. U.A. (1995). Massless particles electromagnetism, and Rieffel induction. *Reviews in Mathematical Physics* 7, 923–958.

Lee, C.W. (1976). A restriction on the topology of Cauchy surfaces in general relativity. *Communications in Mathematical Physics* 51, 157–162.

Lee, J.M., Parker, T.H. (1987). The Yamabe Problem. *Bulletin of the American Mathematical Society* 17, 37–91.

Lee, J., Wald, R.M. (1990). Local symmetries and constraints *Journal of Mathematical Physics* 31, 725–743.

Lesourd, M. (2018). A new singularity theorem for black holes which allows chronology violation in the interior. *Classical and Quantum Gravityt* 35:245003.

Lesourd, M. (2019). Cosmological singularities from high matter density without global topological assumptions. *General Relativity and Gravitation* 51:113.

Li, J., Yu, P. (2015). Construction of Cauchy data of vacuum Einstein field equations evolving to black holes. *Annals of Mathematics* 181, 699–768.

Lieb, E.H., Yngvason, J. (1999). The physics and mathematics of the second law of thermodynamics. *Physics Reports* 310, 1–96.

Lindblad, H., Rodnianski, I. (2003). Global existence for the Einstein vacuum equations in wave coordinates. *Communications in Mathematical Physics* 256, 43–110.

Lindblad, H., Rodnianski, I. (2010). The global stability of Minkowski space-time in harmonic gauge. *Annals of Mathematics* 1401–1477.

Lopes Costa, J. (2010). *On Black Hole Uniqueness Theorems*. DPhil Thesis, Oxford University.

Lovelock, D. (1971). The Einstein tensor and its generalizations. *Journal of Mathematical Physics* 12, 498–501.

Luk, J. (undated). *Introduction to Nonlinear Wave Equations*. `https://www.dpmms.cam.ac.uk/~jl845/NWnotes.pdf`.

Luk, J. (2012). On the local existence for the characteristic initial value problem in general relativity. *International Mathematics Research Notices* 20, 4625–4678.

Luk, J., Oh, S.-J. (2019a). Strong cosmic censorship in spherical symmetry for two-ended asymptotically flat Initial data I: The interior of the black hole region. *Annals of Mathematics* 190, 1–111.

Luk, J., Oh, S.-J. (2019b). Strong cosmic censorship in spherical symmetry for two-ended asymptotically flat Initial data II: The exterior of the black hole region. *Annals of PDE* 5:6.

Luminet, J.P. (1979). Image of a spherical black hole with thin accretion disk. *Astronomy and Astrophysics* 75, 228–235.

Luna, R., Zilhao, M., Cardoso, V., Costa, J.L., Natário, J. (2019). Strong cosmic censorship: The nonlinear story. *Physical Review D* 99:064014.

Manchak, J.B. (2011). What is a physically reasonable spacetime? *Philosophy of Science* 78, 410–420.

Manchak, J.B. (2014). On space-time singularities, holes, and extensions. *Philosophy of Science* 81, 1066–1076.

Manchak, J.B. (2017). On the inextendibility of space-time. *Philosophy of Science* 84, 1215–1225.

Măcuț, I. (2016). *Manifolds*. `http://www.math.ru.nl/~imarcut/index_files/lectures_2016.pdf`.

Mars, M. (2009). Present status of the Penrose inequality. *Classical and Quantum Gravity* 26:193001.

Marsden, J.E., Weinstein, A. (1974). Reduction of symplectic manifolds with symmetry. *Reports on Mathematical Physics* 5, 121–130.

Marsden, J.E., Weinstein, A. (2001). Comments on the history, theory, and applications of symplectic reduction. *Quantization of Singular Symplectic Quotients*, eds. Landsman, N.P., Pflaum, M., Schlichenmaier, M., pp. 1–19 (Birkhäuser).

Martín-Moruno, P., Visser, M. (2017). Classical and semi-classical energy conditions. *Wormholes, Warp Drives and Energy Conditions*, eds. Lobo, F.S.N., pp. 193–213 (Springer).

Masood-ul-Alam, A.K.M. (1992). Uniqueness proof of static charged black holes revisited. *Classical and Quantum Gravity* 9, L53–L55.

Maudlin, T. (2002). Thoroughly muddled Mctaggart: Or, how to abuse gauge freedom to create metaphysical monstrosities. With a response by John Earman. *Philosophers Imprint* 2, no. 4, 1–23.

McFeron, D., Székelyhidi, G. (2012). On the positive mass theorem for manifolds with corners. *Communications in Mathematical Physics* 313, 425–443.

McNamara, J.M. (1978). Instability of black hole inner horizons. *Proceedings of the Royal Society A* 358, 499–517.

McTaggart Ellis McTaggart, J. (1908). The unreality of time. *Mind* 17, 457–474.

Meeks, W.H., Yau, S.T. (1980). Topology of three dimensional manifolds and the embedding problems in minimal surface theory. *Annals of Mathematics* 112, 441–484.

Melrose, R. (1996). Differential analysis on manifolds with corners, in preparation. `http://www-math.mit.edu/~rbm/book.html`.

Meyer, K. (1973). Symmetries and integrals in mechanics. *Dynamical Systems*, ed. Peixoto, M.M., pp. 259–272 (Academic Press).

Minguzzi, E. (2008a). Limit curve theorems in Lorentzian geometry. `arXiv:0712.3942`.

Minguzzi, E. (2008b). Non-imprisonment conditions on spacetime. `arXiv:0712.3949.pdf`.

Minguzzi, E. (2015a). Convex neighborhoods for Lipschitz connections and sprays. *Monatshefte für Mathematik* 177, 569–625.

Minguzzi, E. (2015b). The vacuum conservation theorem. *General Relativity and Gravitation* 47:32.

Minguzzi, E. (2019). Lorentzian causality theory, *Living Reviews in Relativity* 22:3.

Minguzzi, E. (2020). A gravitational collapse singularity theorem consistent with black hole evaporation. *Letters in Mathematical Physics* 110, 2383–2396.

Minguzzi, E., Sánchez, M. (2008). The causal hierarchy of spacetimes. *Recent Developments in Pseudo-Riemannian Geometry*, ed. Alekseevskii, D.A., pp. 299–368 (European Mathematical Society).

Minguzzi, E., Suhr, S. (2019). Some regularity results for Lorentz–Finsler spaces. *Annals of Global Analysis and Geometry* 56, 597–611.

Misner, C.W. (1967). Taub–NUT space as a counterexample to almost anything. *Relativity Theory and Astrophysics I: Relativity and Cosmology*, ed. Ehlers, J., pp. 160–169 (American Mathematical Society).

Moncrief, V., Teitelboim, C. (1972). Momentum constraints as integrability conditions for the hamiltonian constraint in general relativity. *Physical Review D* 6, 966–968.

Moschella, U. (2005). The de Sitter and anti de Sitter sightseeing tour. *Séminaire Poincaré* 1, 1–12.

Müller zum Hagen, H. (1970a). On the analyticity of static vacuum solutions of Einstein's equation. *Mathematical Proceedings of the Cambridge Philosophical Society* 67, 415–421.

Müller zum Hagen, H. (1970b). On the analyticity of stationary vacuum solutions of Einstein's equation. *Mathematical Proceedings of the Cambridge Philosophical Society* 68, 199–201.

Narayan, R., Johnson, M.D., Gammie, F. (2019). The shadow of a spherically accreting black hole. *The Astrophysical Journal Letters* 885:L33.

Nathanail, A., Most, E.R., Rezzolla, L. (2017). Gravitational collapse to a Kerr–Newman black hole. *Monthly Notices of the Royal Astronomical Society* 469, L31–L35.

Naumann, J., Simader, C.G. (2011). Measure and integration on Lipschitz-manifolds. `https://edoc.hu-berlin.de/bitstream/handle/18452/3425/15.pdf?sequence=1`.

Navarro, A., Navarro, J. (2010). Lovelock's Theorem revisited. `arXiv:1005.2386`.

Newman, E., Penrose, R. (1962). An approach to gravitational radiation by a method of spin coefficients. *Journal of Mathematical Physics* 3, 566–578.

Nomizu, K., Ozeki, H. (1961). The existence of complete Riemannian metrics, *Proceedings of the American Mathematical Society* 12, 889–891.

Norton, J.D. (2010). Time really passes. *Journal of Philosophical Studies* 13, 23–34.

Ong, Y.C. (2020). Spacetime singularities and cosmic censorship conjecture: A review with some thoughts. `arXiv:2005.07032`.

Oosterhout, W. van (2019). *Birkhoff's Theorem in General Relativity*. B.Sc. Thesis, Radboud University Nijmegen. `https://annegretburtscher.files.wordpress.com/2019/11/willemvanoosterhout_bscthesis_2019.pdf`.

O' Raifeartaigh, L. (1958). Fermi coordinates. *Proceedings of the Royal Irish Academy. Section A: Mathematical and Physical Sciences* 59, 15–24.

Palomo, F.J., Romero, A. (2006). Certain actual topics on modern Lorentzian geometry. *Handbook of Differential Geometry, Vol. II*, eds. Dillen, F.J.E., Verstraelen, L.C.A., pp. 513–546 (Elsevier).

Paetz, T.T. (2014). Characteristic initial data and smoothness of Scri. II. Asymptotic expansions and construction of conformally smooth data sets. *Journal of Mathematical Physics* 55, 102503.

Parker, T., Taubes, C.H. (1982). On Witten's proof of the positive energy theorem. *Communications in Mathematical Physics* 84, 223–238.

Penrose, R. (1973). Naked singularities. *Annals of the New York Academy of Sciences* 224, 125–134.

Penrose, R. (1974). Singularities in cosmology. *Confrontation of Cosmological Theories with Observational Data*, ed. Longair, M.S., pp. 263–272 (D. Reidel).

Penrose, R. (1979). Singularities and time-asymmetry. *General Relativity: An Einstein Centenary Survey*, eds. Hawking, S.W., Israel, W., pp. 581–638 (Cambridge University Press).

Penrose, R. (1999). The question of cosmic censorship. *Journal of Astrophysics and Astronomy* 20, 233–248.

Penrose, R. Floyd, R.M. (1971). Extraction of rotational energy from a black hole. *Nature Physical Science* 229, 177–179.

Pitts, J. (2014). Change in Hamiltonian general relativity from the lack of a time-like Killing vector field. *Studies in History and Philosophy of Modern Physics* 47, 68–89.

Planck, M. (1926). Über die Begrundung des zweiten Hauptsatzes der Thermodynamik. *Sitzungsberichte der Preussischen Akademie der Wissenschaften, Physikalisch-Mathematische Klasse* 453–463.

Poincaré, H. (1885). Sur les courbes définies par les équations differentielles. *Journal de Mathématiques Pures et Appliquées* 4, 167–244.

Pooley, O. (2013). Substantivalist and relationalist approaches to spacetime. *The Oxford Handbook of Philosophy of Physics*, ed. R. Batterman, chapter 16, (Oxford University Press).

Pooley, O. (2015). Background independence, diffeomorphism invariance, and the meaning of coordinates. *Towards a Theory of Spacetime Theories*, eds. D. Lehmkuhl, G. Schiemann, E. Scholz, pp. 105–144 (Springer).

Pooley, O. (2020). The hole argument. `arXiv:2009.09982`.

Rácz, I. (2015). Constraints as evolutionary systems. *Classical and Quantum Gravity* 33:015014.

Rácz, I., Wald, R.M. (1996). Global extensions of spacetimes describing asymptotic final states of black holes. *Classical and Quantum Gravity* 13, 539–552.

Read, J., Brown, H.R., Lehmkuhl, D. (2018). Two miracles of general relativity. *Studies in History and Philosophy of Modern Physics* 64, 14–25.

Reall, H. (2020). *Black Holes*. Part III Lectures, University of Cambridge. `http://www.damtp.cam.ac.uk/user/hsr1000/black_holes_lectures_2016.pdf`.

Regge, T., Teitelboim, C. (1974). Role of surface integrals in the Hamiltonian formulation of general relativity. *Annals of Physics* 88, 286–318.

Rendall, A.D. (1990). Reduction of the characteristic initial value problem to the Cauchy problem and its applications to the Einstein equations. *Proceedings of the Royal Society of London A* 427, 221–239.

Rendall, A.D. (2005). Theorems on existence and global dynamics for the Einstein equations. *Living Reviews in Relativity* 8:6.

Rovelli, C. (2019). Neither presentism nor eternalism. *Foundations of Physics* 49, 1325–1335.

Sánchez, M. (2005). Causal hierarchy of spacetimes, temporal functions and smoothness of Geroch's splitting. A revision, `arXiv:gr-qc/0411143`.

Sánchez, M. (2007). Recent progress on the notion of global hyperbolicity. `arXiv:gr-qc/0712.1933`.

Sanchis-Gual, N., Degollado, J.C., Montero, P.J., Font, J.A., Herdeiro, C. (2016). Explosion and final state of an unstable Reissner-Nordström black hole. *Physical Review Letters* 116, 141101

Sbierski, J. (2016). On the existence of a maximal Cauchy development for the Einstein equations: A dezornification. *Annales Henri Poincaré* 17, 301–329.

Sbierski, J. (2018a). The C^0-inextendibility of the Schwarzschild spacetime and the spacelike diameter in Lorentzian geometry. *Journal of Differential Geometry* 108, 319–378.

Sbierski, J. (2018b). On the proof of the C^0-inextendibility of the Schwarzschild spacetime. *Journal of Physics: Conference Series* 968:012012.

Schaaf, N. van der (2020). *Diffeology, Groupoids, and Morita Equivalence*. M.Sc Thesis, Radboud University Nijmegen. `https://www.math.ru.nl/~landsman/NestaM.pdf`.

Schoen, R. (1984). Conformal deformation of a Riemannian metric to constant scalar curvature. *Journal of Differential Geometry* 20, 479–495.

Schoen, R. (1989). Variational theory for the total scalar curvature functional for Riemannian metrics and related topics. *Lecture Notes in Mathematics* 1365, 120–154.

Schoen, R. (2009). *Topics in Differential Geometry*. `http://math.stanford.edu/~schoen/trieste2012/`.

Schoen, R., Yau, S.T. (1979). On the proof of the positive mass conjecture in general relativity. *Communications in Mathematical Physics* 65, 45–76.

Schoen, R., Yau, S.T. (1981). Proof of the positive mass theorem. II. *Communications in Mathematical Physics* 79, 231–260.

Schoen, R., Yau, S.T. (1983). The existence of a black hole due to condensation of matter. *Communications in Mathematical Physics* 65, 575–579.

Schmidt, B. (1967). Isometry groups with surface-orthogonal trajectories. *Zeitschrift für Naturforschung* A22, 1351–1355.

Schultz, R. (undated). Manifolds with boundary. University of California at Riverside, notes. `https://math.ucr.edu/~res/math260s10/manwithbdy.pdf`.

Seeley, R.T. (1964). Extension of C^∞ functions defined in a half space. *Proceedings of the American Mathematical Society* 15, 625–626.

Seifert, H.J. (1967). Global connectivity by timelike geodesics, *Zeitschrift für Naturforschung A* 22, 1356–1360.

Senovilla, J.M.M. (1998). Singularity theorems and their consequences. *General Relativity and Gravitation* 30, 701–848. Corrected version: `arXiv:1801.04912`.

Senovilla, J.M.M. (2010). Black holes and trapped surfaces. *AIP Conference Proceedings* 1318, 123–135.

Shoemaker, S. (1969). Time without change. *Journal of Philosophy* 66, 363–381.

Simpson, M., Penrose, R. (1973). Internal instability in a Reissner–Nordström black hole. *International Journal of Theoretical Physics* 7, 183–197.

Smulevici, J. (2017). *Lectures on Lorentzian geometry and hyperbolic pdes*. `https://www.math.u-psud.fr/~smulevic/lgpdes.pdf`.

Stewart, J.M., Friedrich, H. (1982). Numerical relativity. I. The characteristic initial value problem. *Proceedings of the Royal Society of London A* 384, 427–454.

Strien, M. van (2014). On the origins and foundations of Laplacian determinism. *Studies in History and Philosophy of Science Part A* 45, 24–31.

Sudarsky, D., Wald, R.M. (1992). Extrema of mass, stationarity, and staticity, and solutions to the Einstein-Yang-Mills equations. *Physical Review D* 46, 1453–1474.

Sudarsky, D., Wald, R.M. (1993). Mass formulas for stationary Einstein–Yang-Mills black holes and a simple proof of two staticity theorems. *Physical Review D* 47, R5209–5213.

Teitelboim, C. (1973). How commutators of constraints reflect the spacetime structure. *Annals of Physics* 79, 542–557.

Teo, E. (2003). Spherical photon orbits around a Kerr black hole. *General Relativity and Gravitation* 35, 1909–1926.

Thébault, K. (2021). The problem of time. *Routledge Companion to the Philosophy of Physics*, eds. Knox, E., Wilson, A., in press (Routledge).

Thorne, K.S. (1993). Misner space as a prototype for almost any pathology. *Directions in General Relativity, Vol. 1*, eds. Hu, B.L., Ryan, M.P., Vishveshwara, C.V., pp. 333-346 (Cambridge University Press).

Tipler, F.J., Clarke, C.J.S., Ellis, G.F.R. (1980). Singularities and horizons–A review article. *General Relativity and Gravitation: One Hundred Years After the Birth of Albert Einstein*, Vol. 2. ed. Held, A., pp. 97–206 (Plenum Press).

Uffink, J. (2007). Compendium of the foundations of classical statistical physics. *Handbook of the Philosophy of Science. Vol. 2: Philosophy of Physics, Part B*, eds. Butterfield, J., Earman, J., pp. 923–1074 (North-Holland).

Van de Moortel, M. (2020). The breakdown of weak null singularities inside black holes. `arXiv:1912.10890`.

Visser, M. (2006). The Kerr spacetime: A brief introduction. `arXiv:0706.0622`.

Volkov, M.S. (2018). Hairy black holes in the XX-th and XXI-st centuries. *The Fourteenth Marcel Grossmann Meeting*, pp. 1779–1798. `https://doi.org/10.1142/9789813226609_0184`.

Wald, R.M. (1993). Black hole entropy is the Noether charge. *Physical Review D* 48, R3427–R3431.

Wald, R.M. (2001). The thermodynamics of black holes. *Living Reviews in Relativity* 4:6.

Wall, A.C. (2009). Ten proofs of the generalized second law. *Journal of High Energy Physics* JHEP06(2009)021.

Wall, A.C. (2018). A survey of black hole thermodynamics. `arXiv:1804.10610`.

Wallace, D. (2017). Who's afraid of coordinate systems? An essay on representation of spacetime structure. *Studies in History and Philosophy of Modern Physics* 67, 125–136.

Wallace, D. (2018). The case for black hole thermodynamics part I: Phenomenological thermodynamics. *Studies in History and Philosophy of Modern Physics* 64, 52–67.

Wallace, D. (2019). The case for black hole thermodynamics part II: Statistical mechanics. *Studies in History and Philosophy of Modern Physics* 66, 103–117.

Weatherall, J.O. (2018). Regarding the 'hole argument'. *British Journal for the Philosophy of Science* 69, 329–350.

Weinstein, G. [Gilbert] (1996). *N*-Black hole stationary and axially symmetric solutions of the Einstein/Maxwell equations *Communications in Partial Differential Equations* 21, 1389–1430.

Weinstein, G. [Galina] (2021). Demons in black hole thermoydynamics: Bekenstein and Hawking. `arXiv:2102.11209`.

Weyl, H. (1938). Courant and Hilbert on partial differential equations. *Bulletin of the American Mathematical Society* 44, 602–604.

Wieting, T. (2010). Capturing infinity: The Circle Limit series of M.C. Escher. *Reed Magazine* March, 21–29.

Wigner, E.P. (1939). Unitary representations of the inhomogeneous Lorentz group. *Annals of Mathematics* 40, 149–204.

Witten, E. (1981). A new proof of the positive energy theorem. *Communications in Mathematical Physics* 80, 381–402.

Wong, W.-Y. (2009). *On the Uniqueness of Kerr–Newman Black Holes*. PhD Thesis, Princeton University.

Wong, W.-Y. (2013). A comment on the construction of the maximal globally hyperbolic Cauchy development. *Journal of Mathematical Physics* 54, 113511

Zilhão, M., Cardoso, V., Herdeiro, C., Lehner, L., Sperhake, U. (2014). Testing the nonlinear stability of Kerr-Newman black holes. *Physical Review D* 90:124088.

Index

F

G

H

I

J

K

L

M

R

S

This book, dedicated to Roger Penrose, is a second, mathematically oriented course in general relativity. It contains extensive references and occasional excursions in the history and philosophy of gravity, including a relatively lengthy historical introduction. The book is intended for all students of general relativity of any age and orientation who have a background including at least first courses in special and general relativity, differential geometry, and topology. The material is developed in such a way that through the last two chapters the reader may acquire a taste of the modern mathematical study of black holes initiated by Penrose, Hawking, and others, as further influenced by the initial-value or PDE approach to general relativity. Successful readers might be able to begin reading research papers on black holes, especially in mathematical physics and in the philosophy of physics. The chapters are: Historical introduction, General differential geometry, Metric differential geometry, Curvature, Geodesics and causal structure, The singularity theorems of Hawking and Penrose, The Einstein equations, The 3+1 split of space-time, Black holes I: Exact solutions, and Black holes II: General theory. These are followed by two appendices containing background on Lie groups, Lie algebras, & constant curvature, and on Formal PDE theory.

Klaas Landsman (1963) has been a professor of mathematical physics since 2001, initially at the University of Amsterdam and since 2004 at Radboud University, where he is a founding member of the Institute for Mathematics, Astrophysics, and Particle Physics (IMAPP). He was a postdoc at the University of Cambridge from 1989-1997, and a research fellow of the Royal Netherlands Academy of Arts and Sciences (KNAW) from 1997-2002. He has been an elected member of the KNAW since 2019. His previous Open Access books include Foundations of Quantum Theory (2017) and The Challenge of Chance (2016). He also wrote two popular science books in Dutch. In 2020 he won the international FQXi essay contest on Undecidability, Uncomputability, and Unpredictability.

Radboud University

www.ru.nl/radbouduniversitypress

ISBN 9 789083 178929

Cover Art: Edith de Jong