Canonical Gravity and Applications

Cosmology, Black Holes, and Quantum Gravity

MARTIN BOJOWALD

CAMBRIDGE

CAMBRIDGE www.cambridge.org/9780521195751

This page intentionally left blank

CANONICAL GRAVITY AND APPLICATIONS

Canonical methods are a powerful mathematical tool within the field of gravitational research, both theoretical and observational, and have contributed to a number of recent developments in physics. Providing mathematical foundations as well as physical applications, this is the first systematic explanation of canonical methods in gravity. The book discusses the mathematical and geometrical notions underlying canonical tools, highlighting their applications in all aspects of gravitational research, from advanced mathematical foundations to modern applications in cosmology and black-hole physics. The main canonical formulations, including the Arnowitt–Deser–Misner (ADM) formalism and Ashtekar variables, are derived and discussed.

Ideal for both graduate students and researchers, this book provides a link between standard introductions to general relativity and advanced expositions of black hole physics, theoretical cosmology, or quantum gravity.

Martin Bojowald is an Associate Professor at the Institute for Gravitation and the Cosmos, Pennsylvania State University. He pioneered loop quantum cosmology, a field in which his research continues to focus.

CANONICAL GRAVITY AND APPLICATIONS

Cosmology, black holes, quantum gravity

MARTIN BOJOWALD
Institute for Gravitation and the Cosmos
The Pennsylvania State University

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore,
São Paulo, Delhi, Dubai, Tokyo, Mexico City

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521195751

First published 2011

Printed in the United Kingdom at the University Press, Cambridge

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

Library of Congress Cataloguing in Publication data
Bojowald, Martin.
Canonical gravity and applications : cosmology, black holes, and quantum gravity / Martin Bojowald.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-521-19575-1 (hardback)
1. Quantum gravity. 2. General relativity (Physics) 3. Cosmology. I. Title.
QC178.B625 2010
530.14′3 – dc22 2010038771

ISBN 978-0-521-19575-1 Hardback

Contents

1	Introduction	*page* 1
2	Isotropic cosmology: a prelude	4
	2.1 Equations of motion	5
	2.2 Matter parameters	8
	2.3 Energy conditions	10
	2.4 Singularities	11
	2.5 Linear perturbations	12
3	Hamiltonian formulation of general relativity	17
	3.1 Constrained systems	18
	3.2 Geometry of hypersurfaces	40
	3.3 ADM formulation of general relativity	50
	3.4 Initial-value problem	72
	3.5 First-order formulations and Ashtekar variables	82
	3.6 Canonical matter systems	98
4	Model systems and perturbations	113
	4.1 Bianchi models	113
	4.2 Symmetry	129
	4.3 Spherical symmetry	147
	4.4 Linearized gravity	163
5	Global and asymptotic properties	184
	5.1 Geodesic congruences	185
	5.2 Trapped surfaces	198
	5.3 Asymptotic infinity	203
	5.4 Matching of solutions	217
	5.5 Horizons	235
6	Quantum gravity	248
	6.1 Constrained quantization and background independence	249
	6.2 Quantum cosmology	258

6.3 Quantum black holes 271
6.4 The status of canonical quantum gravity 273
Appendix A: Some mathematical methods 274
References 289
Index 300

1
Introduction

Einstein's equation

$$G_{ab} = 8\pi G T_{ab} \tag{1.1}$$

presents a complicated system of non-linear partial differential equations of up to second order for the space-time metric g_{ab}. As a tensorial equation, it determines the structure of space-time in a covariant and coordinate-independent way. Nevertheless, coordinates are often chosen to arrive at specific solutions, and the Einstein tensor is split into its components in the process. In component form, one then notices that some of the equations are of first order only; they do not appear as evolution equations but rather as constraints on the initial values that can be posed for the second-order part of Einstein's equation. Moreover, some components of the metric do not appear as second-order derivatives at all.

Physically, all these properties taken together capture the self-interacting nature of the gravitational field and its intimate relationship with the structure of space-time. Einstein's equation is not to be solved on a given background space-time, its solutions rather determine how space-time itself evolves starting with the structure of an initial spatial manifold. General covariance allows one to express solutions in any coordinate system and to relate solutions based only on different choices of coordinates in consistent ways. Consistency is ensured by properties of the first-order part of the equation, and coordinate redundancy by the different behaviors of metric components. All these properties are thus crucial, but they make the theory rather difficult to analyze and to understand.

Instead of solving Einstein's equation just as one set of coupled partial differential equations, the use of *geometry* provides important additional insights by which much information can be gained in an elegant and systematic way. There is, first, space-time itself which is equipped with a Riemannian structure and thus encodes the gravitational field in a geometrical way. Geometry allows many identifications of observable space-time quantities, and it provides means to understand space-time globally and to arrive at general theorems, for instance regarding singularities. These structures can be analyzed with differential geometry, which is provided in most introductory textbooks on general relativity and will be assumed at least as basic knowledge in this book. (More advanced geometrical topics are provided in the Appendix.) We will be assuming familiarity with the first part of the book by Wald (1984), and use similar notations.

In addition to space-time, also the solution space to Einstein's equation, just like the solution space of any field theory, is equipped with a special kind of geometry: symplectic or Poisson geometry as the basis of canonical methods. General properties of solution spaces regarding gauge freedom, as originally analyzed by Dirac, are best seen in such a setting. In this book, the traditional treatment of systems with constraints following Dirac's classification will be accompanied by a mathematical discussion of geometrical properties of the solution spaces involved. With this combination, a more penetrating view can be developed, showing how natural several of the distinctions made by Dirac are from a mathematical perspective. In gravity, these techniques become especially important for understanding the solutions of Einstein's equation and their relationships to each other and to observables. They provide exactly the systematic tools required to understand the evolution problem and consistency of Einstein's equation and the meaning of the way in which space-time structure is described, but they are certainly not confined to this purpose. Canonical techniques are relevant for many applications, including cosmology of homogeneous models and perturbations around them, and collapse models of matter distributions into black holes. Regarding observational aspects of cosmology, for instance, canonical methods provide systematic tools to derive gauge-invariant observables and their evolution. Finally, canonical methods are important when the theory is to be quantized to obtain quantum gravity.

We will first illustrate the appearance and application of canonical techniques in gravity by the example of isotropic cosmology. What we learn in this context will be applied to general relativity in Chapter 3, in which the main versions of canonical formulations — those due to Arnowitt, Deser and Misner (ADM) (2008) and a reformulation in terms of Ashtekar variables — are derived. At the same time, mathematical techniques of symplectic and Poisson geometry will be developed. Applications at this general level include a discussion of the initial-value problem as well as an exhibition of canonical methods and their results in numerical relativity. Canonical matter systems will also be discussed in this chapter.

Just as one often solves Einstein's equation in a symmetric context, symmetry-reduced models provide interesting applications of the canonical equations. Classes of these models, general issues of symmetry reduction, and perturbations around symmetric models are the topic of Chapter 4. The main cosmological implications of general relativity will be touched upon in the process. From the mathematical side, the general theory of connections and fiber bundles will be developed in this chapter. Spherically symmetric models, then, do not only provide insights about black holes, but also illustrate the symmetry structures behind the canonical formulation of general relativity (in terms of Lie algebroids).

Chapter 5 does not introduce new canonical techniques, but rather, shows how they are interlinked with other, differential geometric methods often used to analyze global properties of solutions of general relativity. These include geodesic congruences, singularity theorems, the structure of horizons, and matching techniques to construct complicated solutions from simpler ones. The class of physical applications in this chapter will mainly be black holes, regarding properties of their horizons as well as models for their formation in gravitational collapse.

Chaper 6 then provides concluding discussions with a brief, non-exhaustive outlook on the application to canonical quantum gravity. This topic would require an entire book for a detailed discussion, and so here we only use the final chapter to provide a self-contained link from the methods developed in the main body of this book to the advanced topic of quantum gravity. Several books exist by now dedicated to the topic of canonical quantum gravity, to which we refer for further studies.

This book grew out of a graduate course on “Advanced Topics in General Relativity” held at Penn State, taking place with the prerequisite of a one-semester introduction to general relativity that normally covers the usual topics up to the Schwarzschild space-time. In addition to extending the understanding of Einstein’s equation, this course has the aim to provide the basis for research careers in the diverse direction of gravitational physics, such as numerical relativity, cosmology and quantum gravity. The material contained in this book is much more than could be covered in a single semester, but it has been included to provide a wider perspective and some extra background material. If the book is used for teaching, choices of preferred topics will have to be made. The extra material is sometimes used for independent studies projects, as happened during the preparation of this book.

I am grateful to a large number of colleagues and students for collaborations and explorations over several years, in particular to Rupam Das, Xihao Deng, Golam Hossain, Mikhail Kagan, George Paily, Juan Reyes, Aureliano Skirzewski, Thomas Strobl, Rakesh Tibrewala and Artur Tsobanjan, with whom I have worked on issues related to the material in this book. Finally, I thank Hans Kastrup for having instilled in me a deep respect for Hamiltonian methods. One of the clearest memories from my days as a student is a homework problem of a classical-mechanics class taught by Hans Kastrup. It was about Hamilton–Jacobi methods, epigraphed with the quote “Put off thy shoes from off thy feet, for the place whereon thou standest is holy ground.”

2

Isotropic cosmology: a prelude

Cosmology presents the simplest dynamical models of space-time by assuming space to be homogeneous and isotropic on large scales. This reduces the line element to Friedmann–Lemaître–Robertson–Walker (FLRW) form:

$$\mathrm{d}s^2 = -N(t)^2\mathrm{d}t^2 + a(t)^2\mathrm{d}\sigma_k^2 \tag{2.1}$$

with the spatial line element

$$\mathrm{d}\sigma_k^2 = \frac{\mathrm{d}r^2}{1-kr^2} + r^2(\mathrm{d}\vartheta^2 + \sin^2\vartheta\,\mathrm{d}\varphi^2) \tag{2.2}$$

of a 3-space of constant curvature. Only this form is compatible with the assumption of spatial isotropy — the existence of a 6-dimensional isometry group acting transitively on spatial slices $t = \mathrm{const}$ and on tangent spaces — as we will derive in detail in Chapter 4.2.1. The only free functions are the *lapse function* $N(t)$ and the *scale factor* $a(t)$, while the constant curvature parameter k can take the values zero (spatial flatness), plus one (positive spatial curvature; 3-sphere) or minus one (negative spatial curvature; hyperbolic space).

Both the lapse function and the scale factor must be non-zero, and can be assumed positive without loss of generality. The lapse function determines the clock-rate by which the coordinate t measures time. It can be absorbed by using cosmological proper time[1] τ defined via $\mathrm{d}\tau = N(t)\mathrm{d}t$, a differential equation for $\tau(t)$. With a positive $N(t)$, $\tau(t) = \int N(t)\mathrm{d}t$ is a monotonic function and can thus be inverted to obtain $t(\tau)$ to be inserted in $a(t)$ in the metric if we want to transform from t to τ.

The scale factor measures the expansion or contraction of space in time. For a spatially flat model, it can be rescaled by a constant which would simply change the spatial coordinates. (For models with non-vanishing spatial curvature, the rescaling freedom of coordinates is conventionally fixed by normalizing k to be ± 1.) However, unlike $N(t)$ it cannot be completely absorbed in coordinates while preserving the isotropic form of the line element. Its relative change such as the *Hubble parameter* $\dot{a}/a$ or relative acceleration parameters thus do have physical meaning. They are subject to the dynamical equations of isotropic cosmological models.

[1] The notion of proper time refers to observers, in the present case to co-moving ones staying at a fixed point in space and passively following the expansion or contraction of the universe.

2.1 Equations of motion

The dynamics of gravity is determined by the Einstein–Hilbert action

$$S_{\mathrm{EH}}[g] = \int \mathrm{d}^4x \left(\frac{1}{16\pi G}\sqrt{-\det g}\, R + \mathcal{L}_{\mathrm{matter}} \right) \tag{2.3}$$

where g_{ab} is the space-time metric, $\mathcal{L}_{\mathrm{matter}}$ a Lagrangian density for matter and $R = g^{ab}R_{ab} = g^{ab}R_{acb}{}^{c}$ the Ricci scalar. We will later verify that this action indeed produces Einstein's equation; see Example 3.7.

2.1.1 *Reduced Lagrangian*

For an isotropic metric (2.1) it is easy to derive the Ricci scalar:

$$R = 6\left(\frac{\ddot{a}}{N^2 a} + \frac{\dot{a}^2}{N^2 a^2} - \frac{\dot{a}\dot{N}}{aN^3} + \frac{k}{a^2} \right). \tag{2.4}$$

With $\det g = -r^4 \sin^2(\vartheta) N(t)^2 a(t)^6/(1-kr^2)$ we then have the *reduced gravitational action*

$$S^{\mathrm{iso}}_{\mathrm{grav}}[a,N] = \frac{3V_0}{8\pi G}\int \mathrm{d}t\, N a^3 \left(\frac{\ddot{a}}{N^2 a} + \frac{\dot{a}^2}{N^2 a^2} - \frac{\dot{a}\dot{N}}{aN^3} + \frac{k}{a^2} \right) \tag{2.5}$$

$$= -\frac{3V_0}{8\pi G}\int \mathrm{d}t \left(\frac{a\dot{a}^2}{N} - kaN \right) \tag{2.6}$$

integrating by parts in the second step. Note that we do not need to integrate over all of space (and in fact cannot always do so in a well-defined way if space is non-compact) because the geometry of our isotropic space-time is the same everywhere for constant t. An arbitrary constant $V_0 := \int \mathrm{d}r\mathrm{d}\vartheta\mathrm{d}\varphi r^2 \sin\vartheta/\sqrt{1-kr^2}$ thus arises after picking a compact integration region. From now on we will be assuming that V_0 equals one, which can always be achieved by picking a suitable region to integrate over. This identifies the reduced gravitational Lagrangian as

$$L^{\mathrm{iso}}_{\mathrm{grav}} = -\frac{3}{8\pi G}\left(\frac{a\dot{a}^2}{N} - kaN \right). \tag{2.7}$$

Note that it does not depend on the time derivative of the lapse function.

In this derivation, we are commuting the two steps involved in the derivation of reduced equations of motion: we do not use the full equations of motion that are obtained from varying the action (as done explicitly in Example 3.7) and then insert a special symmetric form of solutions, but insert this symmetric form, (2.1), into the action and then derive equations of motion from variations. There is no guarantee in general that this is in fact allowed: equations of motion correspond to extrema of the action functional; if the action is restricted before variation, some extrema might be missed. The reduced action may, in some cases, not produce the correct equations of motion. In the case of interest here, however, it is true that one can proceed in this way and we do so because it is simpler. We will come back to this problem (called symmetric criticality) from a more general perspective in Chapter 4.2.2.

2.1.2 Canonical analysis

In the reduced action, our free functions of time are $a(t)$ and $N(t)$, which lead to the canonical variables $(a, p_a; N, p_N)$. Momenta are derived in the usual way as

$$p_a = \frac{\partial L^{\rm iso}_{\rm grav}}{\partial \dot{a}} = -\frac{3}{4\pi G}\frac{a\dot{a}}{N}, \qquad p_N = \frac{\partial L^{\rm iso}_{\rm grav}}{\partial \dot{N}} = 0\,. \tag{2.8}$$

Because the Lagrangian does not depend on $\dot{N}$, the momentum p_N vanishes identically and is not a degree of freedom. Its vanishing rather presents a *primary constraint* on the canonical variables and their dynamics. Constraints of this form are associated with gauge freedom of the action, and $p_N = 0$ corresponds to the freedom of redefining time: as seen from the line element, $N(t)$ can be absorbed in the choice of the coordinate t. It thus cannot be a physical degree of freedom, and is not granted a non-trivial momentum.

Proceeding with the canonical analysis, we derive the *gravitational Hamiltonian*

$$H^{\rm iso}_{\rm grav} = \dot{a}p_a + \dot{N}p_N - L^{\rm iso}_{\rm grav} = -\frac{2\pi G}{3}\frac{Np_a^2}{a} - \frac{3}{8\pi G}kaN\,. \tag{2.9}$$

Or, keeping a general matter contribution with Hamiltonian $H_{\rm matter}$ and our primary constraint, which can be added since it vanishes, we have the *total Hamiltonian*

$$H^{\rm iso}_{\rm total} = H^{\rm iso}_{\rm grav} + H^{\rm iso}_{\rm matter} + \lambda p_N \tag{2.10}$$

where $\lambda(t)$ is an arbitrary function. This Hamiltonian determines evolution by Hamiltonian equations of motion

$$\dot{N} = \frac{\partial H^{\rm iso}_{\rm total}}{\partial p_N} = \lambda \tag{2.11}$$

$$\dot{p}_N = -\frac{\partial H^{\rm iso}_{\rm total}}{\partial N} = \frac{2\pi G}{3}\frac{p_a^2}{a} + \frac{3}{8\pi G}ka - \frac{\partial H^{\rm iso}_{\rm matter}}{\partial N} \tag{2.12}$$

$$\dot{a} = \frac{\partial H^{\rm iso}_{\rm total}}{\partial p_a} = -\frac{4\pi G}{3}\frac{Np_a}{a} \tag{2.13}$$

$$\dot{p}_a = -\frac{\partial H^{\rm iso}_{\rm total}}{\partial a} = -\frac{2\pi G}{3}\frac{Np_a^2}{a^2} + \frac{3}{8\pi G}Nk - \frac{\partial H^{\rm iso}_{\rm matter}}{\partial a}\,. \tag{2.14}$$

The first equation, (2.11), tells us again that $N(t)$ is completely arbitary as a function of time, for $\lambda(t)$ remained free when we added the primary constraint to the Hamiltonian. The second equation, (2.12), implies a *secondary constraint* because $p_N = 0$ must be valid at all times, and thus $\dot{p}_N = 0$, or

$$-\frac{2\pi G}{3}\frac{p_a^2}{a} - \frac{3}{8\pi G}ka + \frac{\partial H^{\rm iso}_{\rm matter}}{\partial N} = 0\,. \tag{2.15}$$

The third equation, (2.13), reproduces the definition (2.8) of the momentum p_a, whose equation of motion (2.14) then provides a second-order evolution equation for a.[2]

2.1.3 Scalar field

This set of equations for the gravitational variables is accompanied by equations for matter degrees of freedom, if present, which can be derived analogously from an explicit matter Hamiltonian. In isotropic cosmology, the only matter source compatible with the exact symmetries is a scalar field φ, which in minimally coupled form has an action

$$S_{\text{scalar}}[\varphi] = -\int d^4x\sqrt{-\det g}\left(\frac{1}{2}g^{\mu\nu}\partial_\mu\varphi\partial_\nu\varphi + V(\varphi)\right). \tag{2.16}$$

(More generally, there can be non-minimal coupling terms to gravity of the form $\frac{1}{2}\xi R\varphi^2$ with the Ricci scalar R. Any other curvature couplings would require a parameter of dimension length, which is not available at the classical level; only quantum corrections could provide extra terms making use of the Planck length $\ell_{\text{P}} = \sqrt{G\hbar}$.) For isotropic metrics and spatially homogeneous φ, this reduces to the Lagrangian

$$L^{\text{iso}}_{\text{scalar}} = \frac{a^3}{2N}\dot{\varphi}^2 - Na^3V(\varphi) \tag{2.17}$$

which we now analyze canonically.

The scalar has a momentum

$$p_\varphi = \frac{\partial L^{\text{iso}}_{\text{scalar}}}{\partial\dot{\varphi}} = \frac{a^3\dot{\varphi}}{N} \tag{2.18}$$

and the Hamiltonian is

$$H^{\text{iso}}_{\text{scalar}}(\varphi, p_\varphi) = \dot{\varphi}p_\varphi - L^{\text{iso}}_{\text{scalar}}(\varphi, p_\varphi) = \frac{Np_\varphi^2}{2a^3} + Na^3V(\varphi). \tag{2.19}$$

Hamiltonian equations of motion are $\dot{\varphi} = \partial H^{\text{iso}}_{\text{scalar}}/\partial p_\varphi = Np_\varphi/a^3$ which reproduces (2.18) and

$$\dot{p}_\varphi = -\frac{\partial H^{\text{iso}}_{\text{scalar}}}{\partial\varphi} = -Na^3V'(\varphi). \tag{2.20}$$

2.1.4 Friedmann equations

In order to bring the equations in more conventional form, we use (2.13) to eliminate p_a in (2.15) and (2.14). In this way we obtain the *Friedmann equation*

$$\left(\frac{\dot{a}}{aN}\right)^2 + \frac{k}{a^2} = \frac{8\pi G}{3}\frac{1}{a^3}\frac{\partial H^{\text{iso}}_{\text{matter}}}{\partial N} \tag{2.21}$$

[2] Had we not chosen to set $V_0 = 1$, the Lagrangian, the momenta, and the Hamiltonian would have remained multiplied with V_0. In all equations of motion, both sides scale in the same way when V_0 is changed; the dynamics is thus independent of the choice of V_0.

and the *Raychaudhuri equation*

$$\frac{(\dot{a}/N)^{\cdot}}{aN} = -\frac{4\pi G}{3}\left(\frac{1}{a^3}\frac{\partial H^{\rm iso}_{\rm matter}}{\partial N} - \frac{1}{Na^2}\frac{\partial H^{\rm iso}_{\rm matter}}{\partial a}\right). \tag{2.22}$$

For the scalar field,[3]

$$\frac{1}{a^3}\frac{\partial H^{\rm iso}_{\rm scalar}}{\partial N} = \frac{p_\varphi^2}{2a^6} + V(\varphi) \tag{2.23}$$

and

$$-\frac{1}{Na^2}\frac{\partial H^{\rm iso}_{\rm scalar}}{\partial a} = 3\left(\frac{p_\varphi^2}{2a^6} - V(\varphi)\right).$$

The first-order Hamiltonian equations of motion for φ and p_φ can be combined to a second-order equation for φ, the Klein–Gordon equation

$$\frac{(\dot{\varphi}/N)^{\cdot}}{N} - 3\frac{\dot{a}}{Na}\frac{\dot{\varphi}}{N} + V'(\varphi) = 0\,. \tag{2.24}$$

2.2 Matter parameters

In a matter Hamiltonian, formulated in canonical variables, any N-dependence arises only from the measure factor $\sqrt{-\det g}$, and thus the Hamiltonian must be proportional to N. For a homogeneous space-time, we then have

$$\frac{\partial H_{\rm matter}}{\partial N} = \frac{1}{N}H_{\rm matter} = E \tag{2.25}$$

as the matter Hamiltonian measured in proper time, or the energy. (Energy is frame-dependent, in the case of isotropic cosmology amounting to a reference to N. We will exhibit the general frame dependence in the full expressions in Chapter 3.6.) Furthermore, we use the spatial volume $V = a^3$ to define the *energy density*[4]

$$\rho := \frac{E}{V} = \frac{H_{\rm matter}}{Na^3} \tag{2.26}$$

and *pressure*

$$P := -\frac{\partial E}{\partial V} = -\frac{1}{3Na^2}\frac{\partial H_{\rm matter}}{\partial a}\,. \tag{2.27}$$

These quantities, unlike E, are independent under rescaling a or changing the time coordinate. (In an isotropic universe, these two quantities completely determine the stress-energy tensor

$$T_{ab} = \rho u_a u_b + P(g_{ab} + u_a u_b) \tag{2.28}$$

[3] All partial derivatives require the other *canonical* variables to be held fixed while taking them since these are the independent variables in Hamiltonian equations of motion. Thus, $\partial p_\varphi/\partial a = 0$ even though p_φ, according to (2.18), appears to depend on a. However, $\partial\dot{\varphi}/\partial a \neq 0$ because $\dot{\varphi}$ is not a canonical variable held fixed for $\partial/\partial a$.

[4] With our choice of $V_0 = 1$, this is the energy in our integration region divided by the volume of the region. Thanks to homogeneity, this ratio must be the energy density everywhere.

in perfect-fluid form, such that $\rho = T_{ab}u^a u^b$ and $P = T_{ab}v^a v^a$ where $u^a = (\partial/\partial\tau)^a$ with $u_a u^a = -1$ is the fluid 4-velocity and v^a is a unit spatial vector satisfying $v^a u_a = 0$ and $v_a v^a = 1$.)

Thus, we rewrite the Friedmann and Raychaudhuri equations (2.21) and (2.22) as

$$\left(\frac{\dot{a}}{aN}\right)^2 + \frac{k}{a^2} = \frac{8\pi G}{3}\rho \tag{2.29}$$

$$\frac{(\dot{a}/N)^{\cdot}}{aN} = -\frac{4\pi G}{3}(\rho + 3P)\,. \tag{2.30}$$

This set of one first- and one second-order differential equation implies, as a consistency condition, the *continuity equation*

$$\frac{\dot{\rho}}{N} + 3\frac{\dot{a}}{Na}(\rho + P) = 0\,. \tag{2.31}$$

One can also derive this equation from the conservation equation of a perfect-fluid stress-energy tensor.

Notice that these equations only refer to observable quantities, which are the scaling-independent matter parameters ρ and P as well as the *Hubble parameter*

$$\mathcal{H} = \frac{\dot{a}}{aN} \tag{2.32}$$

and the *deceleration parameter*

$$q = -\frac{a(\dot{a}/N)\cdot N}{\dot{a}^2}\,. \tag{2.33}$$

There is no dependence on the rescaling of the scale factor in these parameters, nor is there a dependence on the choice of time coordinate. In fact, all time derivatives appear in the invariant proper-time form $\mathrm{d}/\mathrm{d}\tau = N^{-1}\mathrm{d}/\mathrm{d}t$.

Example 2.1 (de Sitter expansion)

If pressure equals the negative energy density, $P = -\rho$, the energy density and thus the Hubble parameter $\mathcal{H}$ must be constant in time by virtue of (2.31). This behavior is realized when matter contributions are dominated by a positive cosmological constant Λ. In proper time, we then have the Friedmann equation $\dot{a} = \mathcal{H}a$, solved by $a = a_0 \exp(\mathcal{H}\tau)$.

Next to proper time, a parameter often used is conformal time with $N = a$, making (2.1) with $k = 0$ conformally equivalent to flat space-time. In this example, the transformation to conformal time is obtained as $\eta(\tau) = \int e^{-\mathcal{H}\tau}\mathrm{d}\tau = -(\mathcal{H}a(\tau))^{-1}$. Thus, the scale factor as a function of conformal time behaves as $a(\eta) = -(\mathcal{H}\eta)^{-1}$. While proper time can take the whole range of real values, conformal time must be negative. (None of these coordinates covers all of de Sitter space with a flat spatial slicing.) A finite conformal-time interval approaching $\eta \to 0$ corresponds to an infinite amount of proper time. The divergence of $a(\eta)$ for $\eta \to 0$ is thus only a coordinate effect but with no physical singularity since no observer, who must experience proper time in the rest frame, can live to experience the divergence. For later use we note the relationships $a''/a = 2/\eta^2 = 2\dot{a}^2 = 2\mathcal{H}_{\mathrm{conf}}^2$ for conformal-time *derivatives (denoted by primes) and the* conformal Hubble parameter $\mathcal{H}_{\text{conf}} = a'/a = \dot{a} \neq \mathcal{H}$.

In order to solve the equations of isotropic cosmology, an equation of state $P(\rho)$ must be known, or matter degrees of freedom subject to additional equations of motion must be specified. In the preceding example, this was the simple relationship $P = -\rho$. More generally, one may assume a linear relationship $P = w\rho$ with a constant equation-of-state parameter w.

Example 2.2 (Perfect fluid)

A perfect fluid satisfies the equation of state $P = w\rho$ with a constant w. For $w = 0$, the fluid is called dust, *and for $w = 1/3$ we have* radiation *(see Chapter 3.6.3). Solving the continuity equation (2.31) implies that*

$$\rho \propto a^{-3(w+1)} . \tag{2.34}$$

For dust, energy density $\rho \propto a^{-3}$ is thus just being diluted as the universe expands, while radiation with $\rho \propto a^{-4}$ has an additional red-shift factor. In proper time, $N = 1$, and for a spatially flat universe, $k = 0$, the Friedmann equation $(\dot{a}/a)^2 \propto a^{-3(w+1)}$ shows that $a(\tau) \propto (\tau - \tau_0)^{2/(3+3w)}$ for $w \neq -1$ and $a(\tau) \propto \exp(\sqrt{8\pi G\Lambda/3}\,\tau)$ for $w = -1$, where the matter contribution is only from a cosmological constant $\Lambda = \rho = -P$. In conformal time, $N = a$, the Friedmann equation reads $(a'/a^2)^2 \propto a^{-3(w+1)}$ and gives $a(\eta) \propto (\eta - \eta_0)^{2/(1+3w)}$ for $w \neq -1/3$.

In the example, we can see the following properties:

1. Deceleration, $q > 0$, is realized for $w > -\frac{1}{3}$, which includes all normal forms of matter.
2. Solutions are in general *singular*:
   (i) a can diverge at finite proper time for $w < -1$.
   (ii) a can vanish at finite proper time for $w > -1$, which includes in particular dust and radiation.

In both cases, the Ricci scalar diverges and the Friedmann equation ceases to provide a well-posed initial-value problem. (For the limiting value of $w = -1$, we have the maximally symmetric, and thus non-singular, de Sitter space-time of Example 2.1.)

2.3 Energy conditions

In order to distinguish classes of general matter sources, those not necessarily characterized by a single parameter such as w, with physically and causally reasonable properties one defines energy conditions which a stress-energy tensor should satisfy:

Weak energy condition, WEC $T_{ab}v^a v^b \geq 0$ must be satisfied for all timelike v^a (which by continuity implies that it is also satisfied for null vector fields). If this is true, the local energy density will be non-negative for any observer.

In an isotropic space-time the stress-energy tensor $T_{ab} = \rho u_a u_b + P(g_{ab} + u_a u_b)$ must be of perfect-fluid form, for which the WEC directly implies that $\rho = T_{ab}u^a u^b \geq 0$, and

$T_{ab}v^a v^b = (\rho + P)(u_a v^a)^2 + P v_a v^a \geq 0$ gives $\rho + P \geq 0$, since v^a can be arbitrarily close to a null vector.

The WEC requires that $w \geq -1$, which still allows singularities.

Dominant energy condition $T_{ab}v^a v^b \geq 0$ and $T^{ab}v_a$ must be non-spacelike for all timelike v^a. Then, the local energy flow is never spacelike and energy dominates other components of the stress-energy tensor: $T^{00} \geq |T^{ab}|$ for all a, b. With $\rho \geq P$, the speed of sound $\mathrm{d}P/\mathrm{d}\rho \leq 1$ is no larger than the speed of light for small ρ and P.

Strong energy condition, SEC $T_{ab}v^a v^b \geq \frac{1}{2}T_b^b v^a v_a$ for all timelike v^a implies timelike convergence: via Einstein's equation, we then have $R_{ab}v^a v^b \geq 0$ and the expansion of a family of timelike geodesics never increases. (More details of geodesic families are provided in Chapter 5.) For $v^a = u^a$ the 4-velocity of a perfect fluid, we have $\rho \geq -\frac{1}{2}(-\rho + 3P)$ and thus $\rho + 3P \geq 0$ which is satisfied for $w \geq -\frac{1}{3}$. The strong energy condition thus rules out accelerated expansion.

For the gravitational force governing an isotropic universe, the energy conditions imply that gravity is always attractive because the sign of $\rho + 3P$ restricts the sign of $\ddot{a}$ or of q via the Raychaudhuri equation (2.22).

2.4 Singularities

Assuming that the SEC is satisfied, acceleration of an isotropic universe is ruled out, and singularities are guaranteed. We have seen this in Example 2.2 for specific perfect fluid solutions, but it can also be shown more generally: we use the Raychaudhuri equation in proper time in the form

$$\dot{\mathcal{H}} = \left(\frac{\dot{a}}{a}\right)^{\cdot} = \frac{\ddot{a}}{a} - \left(\frac{\dot{a}}{a}\right)^2 = -\frac{4\pi G}{3}(\rho + 3P) - \mathcal{H}^2 \leq -\mathcal{H}^2 \tag{2.35}$$

applying also the Friedmann equation; the last inequality follows from the SEC. Thus, the Hubble parameter satisfies

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\frac{1}{\mathcal{H}} \geq 1 \tag{2.36}$$

which for solutions implies that

$$\frac{1}{\mathcal{H}} - \frac{1}{\mathcal{H}_0} \geq \tau - \tau_0\,. \tag{2.37}$$

If we assume initial values to be such that the universe is contracting at $\tau = \tau_0$ and thus $\mathcal{H}_0 = \mathcal{H}(\tau_0) < 0$,

$$\frac{1}{\mathcal{H}} \geq \frac{1}{\mathcal{H}_0} + \tau - \tau_0$$

is positive for $\tau > \tau_0 + 1/|\mathcal{H}_0|$. Since $1/\mathcal{H}$ was initially negative and is growing at least as fast as $1/\mathcal{H}_0 + \tau - \tau_0$, it must reach zero before the time $\tau_1 = \tau_0 + 1/|\mathcal{H}_0|$; $\mathcal{H}$ diverges within the finite proper time interval $[\tau_0, \tau_1]$, and then ρ diverges as per the Friedmann equation.

Any isotropic space-time containing matter that satisfies the SEC is thus valid only for a finite amount of proper time to the future if it is contracting at one time. By the same arguments, one concludes that an isotropic space-time which is expanding once had a valid past for only a finite amount of proper time.

One may want specific matter ingredients to avoid the formation of a singularity, which in an isotropic space-time can only be achieved by a turn-around in $a(t)$, satisfying both $\dot{a} = 0$ and $\ddot{a} > 0$ to guarantee a minimum of a at some time. The first condition can be realized for positive spatial curvature, $k = 1$, and positive energy density $\rho \geq 0$. However, the SEC does not allow the solution $a = \sqrt{3/8\pi G\rho}$ to be a minimum; such a point can only be the maximum of a recollapsing universe. One has to violate positive-energy conditions for a minimum, but even this may not be sufficient.

The arguments in this section only apply to isotropic space-times, and initially there were hopes that non-symmetric matter perturbations may prevent the singularity from a collapse into a single point. However, these hopes were not realized as general singularity theorems showed. We will come back to these questions in Chapter 5. But first we will discuss the structure of equations of motion and the Hamiltonian formulation of general relativity without assuming any space-time symmetries. This analysis will provide general properties of the dynamical systems in gravitational physics, which are useful for many further applications.

2.5 Linear perturbations

In order to test the stability of isotropic models as well as to investigate how inhomogeneous cosmological structures can evolve in an expanding universe, perturbations of Einstein's equation around FLRW models are essential. Linear perturbations $g_{ab} = {}^0g_{ab} + \delta g_{ab}$ of the space-time metric and $\varphi = {}^0\varphi + \delta\varphi$ of matter fields are the first step. In a simple manner, linear perturbations can be introduced by changing the line element (2.1) to

$$\mathrm{d}s^2 = -N(t)^2(1+\phi(x,t))^2\mathrm{d}t^2 + a(t)^2(1-\psi(x,t))^2\mathrm{d}\sigma_k^2 \,. \qquad (2.38)$$

To linear order in the inhomogeneity functions $\phi(x,t)$ and $\psi(x,t)$, which are considered small, the time part of the line element is thus multiplied with $(1+\phi(x,t))^2$ and the space part with $(1-\psi)^2$, rescaling the lapse function and the scale factor in a position-dependent way. These rescalings cannot be absorbed in a redefinition of the coordinates and thus capture physical effects in the metric. Nevertheless, the precise form of position-dependent terms does depend on coordinate choices, and the functions ϕ and ψ introduced in this way are not coordinate independent or scalar; they do not directly correspond to physical observables.

Specializing to a spatially flat background model with a scalar field φ as matter source, and expanding Einstein's equation in the perturbations ϕ and ψ in a line element of the form (2.38) results in the following equations:

$$D^2\psi - 3\mathcal{H}_{\rm conf}\left(\psi' + \mathcal{H}_{\rm conf}\phi\right) = 4\pi G\left({}^0\varphi'\delta\varphi' - ({}^0\varphi')^2\phi + a^2 \left.\frac{{\rm d}V}{{\rm d}\varphi}\right|_{{}^0\varphi}\delta\varphi\right) \tag{2.39}$$

from the time-time components of the Einstein and stress-energy tensor,

$$\partial_a\left(\psi' + \mathcal{H}_{\rm conf}\phi\right) = 4\pi G\, {}^0\varphi'\partial_a\delta\varphi \tag{2.40}$$

from the time-space components,

$$\psi'' + \mathcal{H}_{\rm conf}\left(2\psi' + \phi'\right) + \left(2\mathcal{H}'_{\rm conf} + \mathcal{H}^2_{\rm conf}\right)\phi = 4\pi G\left({}^0\varphi'\delta\varphi' - a^2\left.\frac{{\rm d}V}{{\rm d}\varphi}\right|_{{}^0\varphi}\delta\varphi\right) \tag{2.41}$$

from diagonal spatial components, and

$$D^2(\phi - \psi) = 0 \tag{2.42}$$

from the off-diagonal spatial ones. All indices are raised and lowered using the flat Euclidean metric δ_{ab} on spatial slices, with indices a, b indicating spatial directions. The equations are written in conformal time, which is useful in combination with the simultaneous application of the flat derivative operator ∂_a with Laplacian $D^2 = \delta^{ab}\partial_a\partial_b$. Coefficients in the equations depend on the conformal Hubble parameter $\mathcal{H}_{\rm conf}$ of the background.

The scalar field, now also inhomogeneous and perturbed as $\varphi(x,t) = {}^0\varphi + \delta\varphi(x,t)$, must satisfy

$$\delta\varphi'' + 2\mathcal{H}_{\rm conf}\delta\varphi' - D^2\delta\varphi + a^2\left.\frac{{\rm d}^2V}{{\rm d}\varphi^2}\right|_{{}^0\varphi}\delta\varphi + 2a^2\left.\frac{{\rm d}V}{{\rm d}\varphi}\right|_{{}^0\varphi}\phi - {}^0\varphi'\left(\phi' + 3\psi'\right) = 0 \tag{2.43}$$

as the linearized Klein–Gordon equation. More details about the derivation and some applications of these equations will be given in Chapter 4.4; here, only their structure shall concern us.

Example 2.3 (Poisson equation)

Evaluating Eq. (2.39) on a slowly expanding background, thus ignoring terms including the Hubble parameter, we have the equation

$$a^{-2}D^2\phi = 4\pi G\left(\frac{{}^0\varphi'}{a}\frac{\delta\varphi'}{a} - \frac{({}^0\varphi')^2}{a^2}\phi + \frac{{\rm d}V}{{\rm d}\varphi}\delta\varphi\right).$$

All conformal-time derivatives are divided by the lapse function (the scale factor in conformal time). Using $\delta N/N = \phi$, the first two terms can be identified as the linear perturbation of $\frac{1}{2}N^{-2}({\rm d}\varphi/{\rm d}t)^2$, which is the kinetic energy density of the scalar field (see (2.23)), while *the last term is the linear perturbation of the potential. Thus, the equation*

$$a^{-2}D^2\phi = 4\pi G\delta\rho$$

is the Poisson equation for the Newtonian potential ϕ of a linearized metric.

As we already saw for the exactly homogeneous FLRW models, the set of dynamical equations is overdetermined: there are more equations than unknowns; five equations for three free functions $\phi(x,t)$, $\psi(x,t)$ and $\delta\varphi(x,t)$. Nevertheless, the system is consistent. For instance, using the background equations for the isotropic variables, one can derive the linearized Klein–Gordon equation by taking a time derivative of (2.39) and combining it in a suitable way with the other components of the linearized Einstein equation. Moreover, integrating (2.40) spatially (both sides are gradients and boundary terms can be dropped since the homogeneous modes have been split off) and taking a time derivative produces exactly Eq. (2.41) upon using the background Klein–Gordon equation. Thus, two of the five equations are redundant and we are left with three equations for three functions. (One of them, (2.42), is easily solved by $\phi = \psi$ whenever, as in the scalar-field case, there is no off-diagonal spatial term in the stress-energy tensor. Such matter is called free of anisotropic stress.)

In addition to the consistency, there is the issue of coordinate dependence as already alluded to. Changing coordinates does not leave the form (2.38) for a perturbed metric invariant, even if we change our coordinates by $x^\mu \mapsto x^\mu + \xi^\mu$ with a vector field ξ^μ whose components are considered of the same order as the inhomogeneities so as to keep the linear approximation. Under such a transformation, our line element, for perturbations of a spatially flat model, will still describe perturbations of an isotropic geometry of the same order as before, but in general now takes the form

$$\begin{aligned} \mathrm{d}s^2 = & -(N(t)(1+\phi(x,t)))^2\,\mathrm{d}t^2 + N(t)\partial_a B(x,t)\mathrm{d}t\mathrm{d}x^a \\ & + \left((a(t)(1-\psi(x,t)))^2\,\delta_{ab} + \partial_a\partial_b E(x,t)\right)\mathrm{d}x^a\mathrm{d}x^b \end{aligned} \tag{2.44}$$

with two new perturbations $E(x,t)$ and $B(x,t)$. (This is in fact the most general linear perturbation of (2.1) by spatial scalars; vectorial and tensorial perturbations will be introduced in Chapter 4.4.) As long as the coordinate change is linear in perturbations, any such system of coordinates would be equally good.

As discussed in detail in Chapter 4, if one looks at the transformation properties of all components in (2.44), one can see that the combinations

$$\Psi = \psi - \mathcal{H}_{\mathrm{conf}}(B - E') \tag{2.45}$$

and

$$\Phi = \phi + \left(B - E'\right)' + \mathcal{H}_{\mathrm{conf}}(B - E') \tag{2.46}$$

(called Bardeen variables after Bardeen (1980)) are left invariant. Similarly, there is a coordinate-independent combination of perturbations involving the scalar field. Equations of motion derived for the perturbed line element (2.44) involve all components ϕ, ψ, E

and B, but the terms combine in such a way that only Ψ and Φ appear, together with the invariant matter perturbation. In fact, they are merely obtained by replacing ϕ, ψ and $\delta\varphi$ in (2.40)–(2.42) by their coordinate-independent counterparts.

All this, of course, is as it should be. Equations of motion must form a consistent set, and they must not depend on what space-time coordinates are used to formulate them. But mathematically, these properties are certainly not obvious for an arbitrary set of equations. The equations of general relativity ensure that consistency and coordinate independence are realized by the fact that the Einstein tensor is conserved and is indeed a tensor transforming appropriately under coordinate changes. These two crucial properties appear somewhat unrelated at the level of equations of motion, or of Lagrangians. But as we will see in the detailed analysis of canonical gravity to follow, they are closely interrelated. The conservation law implies the existence of constraints, such as those seen in the beginning of this chapter, and the constraints are the generators of space-time symmetries. By fulfilling certain algebraic properties, they guarantee consistency. In addition to revealing these insights, a Hamiltonian formulation has many extra advantages in an analysis of the structure and implications of dynamical equations.

Exercises

2.1 Consider the line element

$$\mathrm{d}s^2 = g_{\mu\nu}\mathrm{d}x^\mu \mathrm{d}x^\nu = -N(t)^2\mathrm{d}t^2 + a(t)^2(\mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2)$$

with arbitrary positive functions $N(t)$ and $a(t)$.

(i) Show that the Ricci scalar is given by

$$R = 6\left(\frac{\ddot{a}}{aN^2} + \frac{\dot{a}^2}{a^2N^2} - \frac{\dot{N}}{N^3}\frac{\dot{a}}{a}\right).$$

(ii) Derive Einstein's equation for an isotropic perfect fluid.

2.2 Compute the Ricci tensor and scalar for the 3-dimensional line element

$$\mathrm{d}s^2 = h_{ab}\mathrm{d}x^a\mathrm{d}x^b = a^2\left(\frac{\mathrm{d}r^2}{1-kr^2} + r^2(\mathrm{d}\vartheta^2 + \sin^2\vartheta\,\mathrm{d}\varphi^2)\right)$$

with constant k and $a \neq 0$, and verify that $R_{ab} = 2kh_{ab}/a^2$ and $R = 6k/a^2$.

2.3 Show that there is a solution (the so-called *Einstein static universe*) to the Friedmann equation in the presence of a positive cosmological constant for which the space-time line element $\mathrm{d}s^2 = -\mathrm{d}\tau^2 + a^2\mathrm{d}\sigma_1^2$ has a spatial part given by the 3-sphere line element $a^2\sigma_1^2$ with *constant* radius a. Relate the 3-sphere line element to the spatial part of an isotropic model with positive spatial curvature, and find conditions for the scale factor to be constant.

2.4 (i) Use the Friedmann and Raychaudhuri equations to show that any matter system with energy density ρ and pressure P satisfies the *continuity equation* $\dot{\rho} + 3\mathcal{H}(\rho + P) = 0$.

(ii) If $P = w\rho$ with constant w, show that $\rho a^{3(w+1)}$ is constant in time.

2.5 Let matter be given by a scalar field φ satisfying the Klein–Gordon equation

$$g^{\mu\nu}\nabla_\mu\nabla_\nu\varphi - \frac{\mathrm{d}V(\varphi)}{\mathrm{d}\varphi} = 0$$

with potential $V(\varphi)$. Show that a homogeneous field φ in an isotropic space-time with line element $\mathrm{d}s^2 = -N(t)^2\mathrm{d}t^2 + a(t)^2(\mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2)$ satisfies

$$\frac{(\dot{\varphi}/N)^{\dot{}}}{N} - 3\frac{\dot{a}\dot{\varphi}}{aN^2} + \frac{\mathrm{d}V(\varphi)}{\mathrm{d}\varphi} = 0$$

by all three following methods:

(i) specializing the Klein–Gordon equation to an isotropic metric,

(ii) using the Lagrangian

$$L_{\mathrm{scalar}} = -\int \mathrm{d}^3x\sqrt{-\det g}\left(\tfrac{1}{2}g^{\mu\nu}\nabla_\mu\varphi\nabla_\nu\varphi + V(\varphi)\right)$$

and deriving its equations of motion for homogeneous φ and isotropic $g_{\mu\nu}$, and

(iii) transforming from L_{scalar} to the Hamiltonian $H_{\mathrm{scalar}} = \int \mathrm{d}^3x(\dot{\varphi}p_\varphi - L_{\mathrm{scalar}})$ and computing its isotropic Hamiltonian equations of motion.

2.6 Verify explicitly that the scalar of the preceding problem satisfies the continuity equation by computing its energy density and pressure from the Hamiltonian.

2.7 (i) Start with a flat isotropic line element

$$\mathrm{d}s^2 = -N(t)^2\mathrm{d}t^2 + a(t)^2(\mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2)$$

and compute its change, given by $\mathcal{L}_\xi g_{\mu\nu}$, under an infinitesimal coordinate transformation generated by a vector field $\xi^\mu(t,x,y,z)\nabla_\mu$.

(ii) Specialize the change $\delta\mathrm{d}s^2 = \mathcal{L}_\xi g_{\mu\nu}\mathrm{d}x^\mu\mathrm{d}x^\nu$ from (i) to a purely temporal vector field

$$\xi^\mu\nabla_\mu = \xi^t(t,x,y,z)\nabla_t$$

and compare with the change obtained by replacing t with $t + \xi^t(t,x,y,z)$ in $a(t)$ and $N(t)\mathrm{d}t$ and expanding to first order in ξ^t.

(iii) Show that the Bardeen variables Ψ and Φ are invariant under the coordinate transformation of part (ii).

3

Hamiltonian formulation of general relativity

As we will see throughout this book, Hamiltonian formulations provide important insights, especially for gauge theories such as general relativity with its underlying symmetry principle of general covariance. Canonical structures play a role for a general analysis of the systems of dynamical equations encountered in this setting, for the issue of observables, for the specific types of equation as they occur in cosmology or the physics of black holes, for a numerical investigation of solutions, and, last but not least, for diverse sets of issues forming the basis of quantum gravity.

Several different Hamiltonian formulations of general relativity exist. In his comprehensive analysis, Dirac (1969), based on Dirac (1958a) and Dirac (1958b) and in parallel with Anderson and Bergmann (1951), developed much of the general framework of constrained systems as they are realized for gauge theories. (Earlier versions of Hamiltonian equations for gravity were developed by Pirani and Schild (1950) and Pirani et al. (1953). In many of these papers, the canonical analysis is presented as a mere prelude to canonical quantization. It is now clear that quantum gravity entails much more, as indicated in Chapter 6, but also that a Hamiltonian formulation of gravity has its own merits for classical purposes.) The most widely used canonical formulation in metric variables is named after Arnowitt, Deser and Misner (Arnowitt et al. (1962)) who first undertook the lengthy derivations in coordinate-independent form. To date, there are several other canonical formulations, such as one based on connections introduced by Ashtekar (1987), which serve different purposes in cosmology, numerical relativity or quantum gravity.

In all cases, to set up a canonical formulation one introduces momenta for time derivatives of the fields. Any such procedure must be a departure from obvious manifest covariance because one initially refers to a choice of time, and then replaces only time derivatives by momenta in the Hamiltonian, leaving spatial derivatives unchanged. Accordingly, canonical equations of motion are formulated for spatial tensors rather than space-time tensors. Introducing momenta after performing a space-time coordinate transformation would in general result in a different set of canonical variables, and so the setting does not have a direct action of space-time diffeomorphisms on all its configurations.

Although the space-time symmetry is no longer manifest and not obvious from the equations, it must still be present; after all, one is just reformulating the classical theory.

Symmetries in such a context can usefully be analyzed by Hamiltonian methods, which then provides crucial insights for the full framework irrespective of whether it is formulated canonically. The mathematical basis of Hamiltonian methods is provided by symplectic and Poisson geometry.

A thorough presentation requires prerequisites of the general theory of constrained systems and of geometrical concepts for hypersurfaces, related to space-time curvature decompositions. This will be developed first in this chapter. More on the mathematical background material of Poisson geometry and tensor densities, which will become important in later stages, is collected in the Appendix.

3.1 Constrained systems

For the linearized cosmological perturbation equations (2.39)–(2.42), we can already notice that there are variables and types of equation of different forms. Only the perturbation $\psi(x,t)$, appearing in the spatial part of the metric (as well as $E(x,t)$ in the fully perturbed setting (2.44)) enters with second-order derivatives in time. For the perturbation $\phi(x,t)$ of the lapse function (or $B(x,t)$ in (2.44)), only up to first-order derivatives in time are to be taken. Moreover, we can see that, concerning time derivatives, there are first- as well as second-order differential equations.

A closer look at the components of the Einstein tensor reveals that these are general properties realized not just for perturbations: Einstein's tensorial equation $G_{ab} = 8\pi G T_{ab}$, when split into components, contains equations of different types. As a whole, the system of partial differential equations is of second order, and thus an initial-value formulation (whose details can be found later in this chapter) would have to pose the values of fields and their first-order time derivatives. But the time components[1] G^0_0 and G^0_a of the Einstein tensor, unlike purely spatial components, do not contain second-order time derivatives.

There is a simple argument for this using the contracted Bianchi identity $\nabla_a G^a_b = 0$. (When matter is present, the same identity holds for $G^a_b - 8\pi G T^a_b$.) If we write this out and solve for the time derivative, we obtain

$$\partial_0 G^0_\mu = -\partial_a G^a_\mu - \Gamma^\nu_{\nu\kappa} G^\kappa_\mu + \Gamma^\kappa_{\nu\mu} G^\nu_\kappa \,. \tag{3.1}$$

The right-hand side clearly contains time derivatives of, at most, second order since ∂_a is only spatial, which means that G^0_μ on the left-hand side can, at most, be of first order in time derivatives. Those components of the Einstein tensor play the role of constraints on the initial values of second-order equations; the equations $G^0_\mu = 8\pi G T^0_\mu$ relate initial values of fields instead of determining how fields evolve. Another important property then follows from (3.1): the constraints are preserved in time; their time derivative automatically vanishes if the spatial part of Einstein's equation is satisfied and if the constraints hold at one time, making the right-hand side of (3.1) zero.

[1] From now on, we will use $a, b, \ldots$ as abstract tangent-space indices irrespective of the dimension. Whenever it seems advisable to distinguish between space-time and spatial tensors in order to avoid confusion, we will use Greek letters $\mu, \nu, \ldots$ for space-time tensors, and reserve Latin ones for spatial tensors.

Checking the orders of derivatives directly in components of the Einstein tensor will reveal a second property; the calculation is thus useful despite being somewhat cumbersome. We start with the Ricci tensor

$$R_{\mu\nu} = \partial_\kappa \Gamma^\kappa_{\mu\nu} - \partial_\mu \Gamma^\kappa_{\kappa\nu} + \Gamma^\kappa_{\mu\nu}\Gamma^\lambda_{\kappa\lambda} - \Gamma^\kappa_{\lambda\nu}\Gamma^\lambda_{\kappa\mu} \,. \tag{3.2}$$

Second-order time derivatives can come only from the first two terms, since the Christoffel symbols

$$\Gamma^\kappa_{\mu\nu} = \frac{1}{2} g^{\kappa\lambda} \left(\partial_\mu g_{\lambda\nu} + \partial_\nu g_{\mu\lambda} - \partial_\lambda g_{\mu\nu} \right)$$

are of first order in derivatives of the metric. The second term in (3.2) is easier to deal with, since $\Gamma^\kappa_{\kappa\nu} = \frac{1}{2} g^{\kappa\lambda} \partial_\nu g_{\kappa\lambda}$ has only one term. Thus, $\partial_\mu \Gamma^\kappa_{\kappa\nu}$ clearly can acquire second-order time derivatives only for $\mu = 0 = \nu$. In this case, up to terms of lower derivatives in time indicated by the dots, we have

$$\partial_0 \Gamma^\kappa_{\kappa 0} = \frac{1}{2} g^{\kappa\lambda} \partial_0^2 g_{\kappa\lambda} + \cdots$$

Extracting terms of highest time-derivative order in $\partial_\kappa \Gamma^\kappa_{\mu\nu}$ requires that $\kappa = 0$ and leads to

$$\partial_\kappa \Gamma^\kappa_{\mu\nu} = \frac{1}{2} g^{0\lambda} \partial_0 \left(\partial_\mu g_{\lambda\nu} + \partial_\nu g_{\mu\lambda} \right) - \frac{1}{2} g^{00} \partial_0^2 g_{\mu\nu} + \cdots$$

where we split off the second term, since it appears in the same form for all μ and ν. For the different options of μ and ν, we read off

$$\partial_\kappa \Gamma^\kappa_{00} = g^{0\lambda} \partial_0^2 g_{0\lambda} - \frac{1}{2} g^{00} \partial_0^2 g_{00} + \cdots$$

for $\mu = 0 = \nu$,

$$\partial_\kappa \Gamma^\kappa_{0a} = \frac{1}{2} g^{0\lambda} \partial_0^2 g_{\lambda a} - \frac{1}{2} g^{00} \partial_0^2 g_{0a} + \cdots = \frac{1}{2} g^{0b} \partial_0^2 g_{ab} + \cdots$$

for $\mu = 0$ and a spatial $\nu = a$. When both μ and ν are spatial, only the last term

$$\partial_\kappa \Gamma^\kappa_{ab} = -\frac{1}{2} g^{00} \partial_0^2 g_{ab} + \cdots$$

contributes a second-order time derivative.

For the Ricci tensor, this implies that

$$\begin{aligned}
R_{00} &= g^{0\lambda} \partial_0^2 g_{0\lambda} - \frac{1}{2} g^{00} \partial_0^2 g_{00} - \frac{1}{2} g^{\kappa\lambda} \partial_0^2 g_{\kappa\lambda} + \cdots = -\frac{1}{2} g^{ab} \partial_0^2 g_{ab} + \cdots \\
R_{0a} &= \frac{1}{2} g^{0b} \partial_0^2 g_{ab} + \cdots \\
R_{ab} &= -\frac{1}{2} g^{00} \partial_0^2 g_{ab} + \cdots
\end{aligned}$$

At this stage, we can already confirm that a crucial property seen in cosmological models is realized in general: only spatial components g_{ab} of the metric appear with their second-order time derivatives. The other components, g_{00} which plays the role of the lapse function seen earlier, and g_{0a} appear only with lower-order time derivatives; they do not play the same dynamical role as g_{ab} does.

From the Ricci tensor, we obtain the second-order time derivative part of the Ricci scalar,

$$R = (g^{0a}g^{0b} - g^{00}g^{ab})\partial_0^2 g_{ab} + \cdots$$

and, combined, the Einstein tensor $G_{\mu\nu} = R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}R$ with components

$$\begin{aligned}
G_{00} &= -\frac{1}{2}\left(g^{ab} + g_{00}(g^{0a}g^{0b} - g^{00}g^{ab})\right)\partial_0^2 g_{ab} + \cdots \\
G_{0a} &= \frac{1}{2}g^{0b}\partial_0^2 g_{ab} - \frac{1}{2}g_{0a}(g^{0c}g^{0d} - g^{00}g^{cd})\partial_0^2 g_{cd} + \cdots \\
G_{ab} &= -\frac{1}{2}g^{00}\partial_0^2 g_{ab} - \frac{1}{2}g_{ab}(g^{0c}g^{0d} - g^{00}g^{cd})\partial_0^2 g_{cd} + \cdots
\end{aligned}$$

If we finally compute the components G^0_ν of the Einstein tensor with mixed index positions as they feature in the contracted Bianchi identity, we have the combinations

$$\begin{aligned}
G^0_0 &= g^{00}G_{00} + g^{0a}G_{0a} \\
&= -\frac{1}{2}(g^{00}g^{ab} - g^{0a}g^{0b})(1 - g^{00}g_{00} - g^{0c}g_{0c})\partial_0^2 g_{ab} + \cdots \\
G^0_a &= g^{00}G_{0a} + g^{0b}G_{ab} \\
&= \frac{1}{2}(g^{00}g^{cd} - g^{0c}g^{0d})(g^{00}g_{0a} + g^{0b}g_{ab})\partial_0^2 g_{cd} + \cdots
\end{aligned}$$

Writing down the identity $g^{\mu\nu}g_{\nu\lambda} = \delta^\mu_\lambda$ in components, one easily sees that $1 - g^{00}g_{00} - g^{0c}g_{0c}$ and $g^{00}g_{0a} + g^{0b}g_{ab}$ vanish; thus, G^0_0 and G^0_a do not contain any second-order time derivatives at all (while the spatial G^a_b do). The corresponding parts of Einstein's equation are of lower order, as we have already observed for cosmological perturbation equations as well as by the general arguments involving the Bianchi identity. These equations, being of lower order in time derivatives than the whole system, provide constraints on the initial values while the spatial components determine the evolution.

This splitting of the equations is important for several reasons. The presence of constraints shows not only that initial values cannot be chosen arbitrarily, but also that there are underlying symmetries. Constraints of a certain type called first class, as they are realized in general relativity, generate gauge transformations, quite analogously to time translations generated by energy in classical mechanics. Classically, the gauge transformations of general relativity are equivalent to coordinate changes. Gauge invariance thus implies the general covariance under coordinate changes.

Constrained systems, sometimes called singular systems, appear whenever a theory has gauge symmetries. General relativity is one example of a more special class of generally

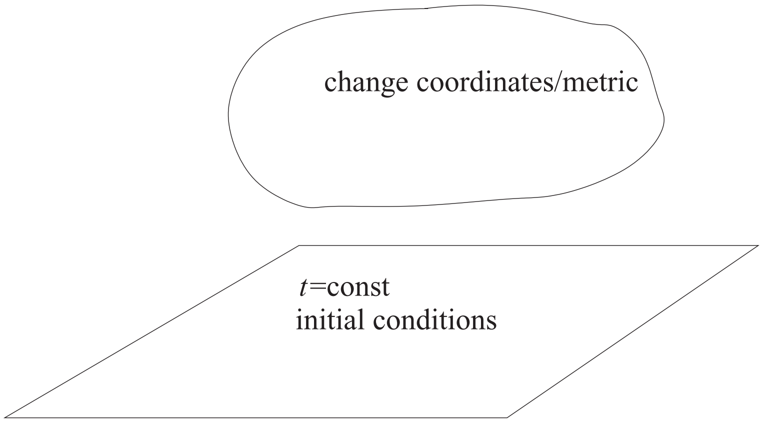

Fig. 3.1 The hole argument: initial values are specified on a spacelike surface in a way that must lead to complete solutions in a deterministic theory. However, one can change coordinates arbitrarily in any region not intersecting the initial data surface, such that a formally different solution is obtained for the same initial values. Physical observables in a deterministic theory must uniquely follow from the initial values. When coordinate changes lead to different representations of solutions evolving from the same initial values, determinism is ensured only if observables are independent of the choice of coordinates.

covariant theories in which the local symmetries are given by coordinate transformations. Since time plays an important role in canonical systems, with a Hamiltonian that generates translations in time — or evolution — but is now, in a covariant setting, subject to gauge transformations, several subtleties arise. This feature lies at the heart of the geometrical understanding of general relativity, and it is the origin of several characteristic and hard problems to be addressed in numerical relativity and quantum gravity.

In practice, the presence of constraints means that the formulation of a theory in terms of fields on a space-time has redundancy. Even though only the geometry is physically relevant, specific, coordinate-dependent values of fields such as the space-time metric at single points are used in any field theoretic setup. Coordinate transformations exist that relate solutions that formally appear different when represented as fields, but that evolve from the same initial values, for instance when the coordinate transformation only affects a region not intersecting the initial data surface, as represented in Fig. 3.1. A deterministic theory, however, cannot allow different solutions to evolve out of the same initial values. Solutions with the same initial values but different field values in a future region must be identified and considered as two different representations of the same physical configuration.

The number of distinguishable physical solutions is thus smaller than one would expect just from the number of initial values required for a set of second-order partial differential equations for a given number of physical fields. Additional restrictions on the initial conditions must exist, which do not take the form of equations of motion but of constraints. This is why the constraints must enter: functionals on the phase space that do not provide equations for time derivatives of canonical variables but rather, non-trivial relations between them. They imply conditions to be satisfied by suitable initial values, but also show how different representations of the same physical solution must be identified.

While the invariance under coordinate transformations is well known and present already in the Lagrangian formulation of general relativity, the canonical formulation involving constraints has several advantages. In this way, the structure of space-time can be analyzed in terms of the algebra of constraints undertaking Poisson brackets, without reference to coordinates. This algebraic viewpoint is important, for instance, in approaches to quantum gravity, where a continuum manifold with coordinates may not be available but instead be replaced by new structures. In terms of quantized constraints, one would still be able to analyze the underlying quantum geometry of space-time. But already, at the classical level, a canonical formulation has several important features, such as the implementation of gauge choices which is easier and more physical to discuss in terms of space-time fields rather than in terms of coordinates.

3.1.1 Lagrangian formulation

Formally, constraints arise from variations of the action just as equations of motion do. At this level, both types of equation are at the same footing, the only difference being that equations of motion are of higher order in time derivatives compared with constraints. One can see how constraints arise directly by following the usual procedure of determining the dynamics from an action principle. On the correct dynamical solutions, the action must be stationary. By the usual variational techniques, applied to a generic first-order action of the form

$$S[q^i(t)] = \int L(q^i, \dot{q}^i)\mathrm{d}t \tag{3.3}$$

for a system of n configuration degrees of freedom q^i, $i = 1, \ldots, n$, one derives the Euler–Lagrange equations

$$-\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{q}^i} + \frac{\partial L}{\partial q^i} = -\frac{\partial^2 L}{\partial \dot{q}^i \partial \dot{q}^j}\ddot{q}^j - \frac{\partial^2 L}{\partial \dot{q}^i \partial q^j}\dot{q}^j + \frac{\partial L}{\partial q^i} = 0\,. \tag{3.4}$$

Clearly, the equations of motion in this case of an action depending on configuration variables and their first-order time derivatives are of second order. But we have a complete set of second-order equations of motion, in general coupled for all the n variables q^i, only if the matrix

$$W_{ij} := \frac{\partial^2 L}{\partial \dot{q}^i \partial \dot{q}^j} \tag{3.5}$$

which multiplies the second-order derivatives of q^i in (3.4) is non-degenerate. If this is so, one can invert W_{ij} and obtain explicit equations of motion

$$\ddot{q}^i = (W^{-1})^{ij}\left(-\frac{\partial^2 L}{\partial \dot{q}^j \partial q^k}\dot{q}^k + \frac{\partial L}{\partial q^j}\right)$$

instead of implicit ones. If W_{ij} is degenerate, on the other hand, the Euler–Lagrange equations (3.4) tell us that q^i and $\dot{q}^i$ must always be such that the vector

$$V_i(q^j, \dot{q}^k) := -\frac{\partial^2 L}{\partial \dot{q}^i \partial q^j} \dot{q}^j + \frac{\partial L}{\partial q^i}, \tag{3.6}$$

required by the variational equations to equal $W_{ij}\ddot{q}^j$, is in the image of $W_{ij} \colon \mathbb{R}^n \to \mathbb{R}^n$, seen as a linear mapping between vector spaces. If W_{ij} is not invertible, the image under this mapping has a non-vanishing co-dimension of the same size as the kernel, and cannot be the full space. The V_i in (3.6) must then lie in a subspace of dimension less than n and cannot be linearly independent, imposing non-trivial restrictions on the fields and the initial values one is allowed to choose for them.

As a symmetric matrix, W_{ij} has $m = n - \operatorname{rank} W$ null-eigenvectors Y_s^i, $s = 1, \ldots, m$, for which $Y_s^i W_{ij} = 0$. Multiplying the Euler–Lagrange equations with the matrix Y_s^i from the left implies that

$$\phi_s(q^i, \dot{q}^i) := Y_s^i \left(-\frac{\partial^2 L}{\partial \dot{q}^i \partial q^j} \dot{q}^j + \frac{\partial L}{\partial q^i} \right) = Y_s^i W_{ij} \ddot{q}^j = 0\,. \tag{3.7}$$

These functionals are constraints to be imposed on q^i and $\dot{q}^i$, and, in particular, on their initial values to be evolved by the second-order part of the equations. As this general discussion shows, equations of motion as well as the constraints are contained in the set of Euler–Lagrange equations.

Example 3.1 (Constraints in isotropic models)
For isotropic cosmology, we have the degrees of freedom $q^i = (N, a)$ *and the symmetric matrix*

$$\left(\frac{\partial^2 L_{\rm grav}^{\rm iso}}{\partial \dot{q}^i \partial \dot{q}^j} \right) = \begin{pmatrix} \frac{\partial^2 L_{\rm grav}^{\rm iso}}{\partial \dot{N}^2} & \frac{\partial^2 L_{\rm grav}^{\rm iso}}{\partial \dot{N} \partial \dot{a}} \\ \frac{\partial^2 L_{\rm grav}^{\rm iso}}{\partial \dot{N} \partial \dot{a}} & \frac{\partial^2 L_{\rm grav}^{\rm iso}}{\partial \dot{a}^2} \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & -\frac{3}{4\pi G} \frac{a}{N} \end{pmatrix}$$

is degenerate of rank one. Its single null-eigenvector $Y = (1, 0)$ *gives rise to the constraint* $\phi = \partial L_{\rm grav}^{\rm iso} / \partial N = C_{\rm grav}^{\rm iso}$ *which we already saw in (2.15).*

3.1.2 Hamiltonian formulation

Like the Lagrangian formulation, a Hamiltonian one directly shows in its general setup how constraints may arise. The usual definition

$$p_i(q^j, \dot{q}^k) = \frac{\partial L}{\partial \dot{q}^i} \tag{3.8}$$

of momenta results in n independent variables only if the matrix $W_{ij} = \partial p_i / \partial \dot{q}^j$, which is identical to (3.5), is invertible such that one can at least locally solve for the $\dot{q}^i$. Otherwise, the map $(q^i, \dot{q}^i) \mapsto (q^i, p_j(q, \dot{q}))$, which in the unconstrained case is a one-to-one

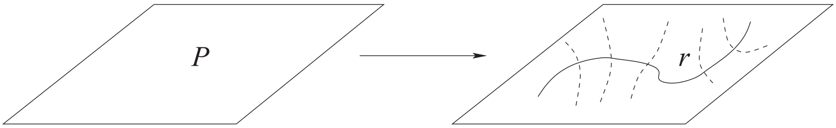

Fig. 3.2 The unconstrained phase space P with coordinates $(q^i, \dot{q}^i)$ is mapped to the primary constraint surface $r\colon \psi_s = 0$ of all points obtained as $(q^i, p_j(q, \dot{q}))$. Dashed lines indicate fibers along which q^i and p_j are fixed but some $\dot{q}^i$ (and the ψ_s) vary.

transformation of variables on the phase space, reduces the dimension as illustrated in Fig. 3.2. The image of this transformation can be characterized by the vanishing of *primary constraints* $\psi_s(q^i, p^j) = 0$, a set of phase-space functions that provides a representation of the image surface. (The ψ_s can be considered as coordinates transversal to the image of $(q^i, \dot{q}^i) \mapsto (q^i, p_j(q, \dot{q}))$.) Locally, the unconstrained phase space of all $(q^i, \dot{q}^i)$ has a complete set of coordinates given by (q^i, p_j, ψ_s). Globally, however, explicit expressions for the ψ_s may be difficult to find. Fortunately, in most cases they follow rather easily from an action just as the momenta do. For instance, if there is a variable, say q^1, whose time derivative does not appear in the action (or appears only via boundary terms such as $\dot{N}$ in isotropic cosmology) $p_1 = \partial L/\partial \dot{q}^1 = 0$ arises immediately as one of the primary constraints.

3.1.2.1 Hamiltonian equations

One may worry that there are obstructions to defining a Hamiltonian in a constrained system because the Legendre transformation

$$H = \dot{q}^i p_i(q, \dot{q}) - L(q, \dot{q}) \tag{3.9}$$

as it stands refers to the time derivatives of $\dot{q}^i$ rather than only to momenta p_j. If the relations (3.8) cannot be inverted to replace all $\dot{q}^i$ in (3.9) by p_j, no phase-space Hamiltonian as a function $H(q, p)$ would exist.

Despite the non-invertibility in the constrained case, H is always a well-defined functional of q^i and p_j. To ensure this, we must show that the value of H as given on the right-hand side of (3.9) does not change when the $\dot{q}^i$ vary while keeping the p_j fixed. Such variations are certainly possible since there are fewer independent p_j than $\dot{q}^i$ in the presence of primary constraints. Using the mapping from $\dot{q}^i$ to p_j, the original phase space of positions and velocities is fibered by submanifolds consisting of all points that are mapped to the same values of (q^i, p_j). For a function to be well defined on (q^i, p_j), we have to make sure that it is constant along those fibers, or that its variation depends only on the changes of q^i and p_j but not on those of $\dot{q}^k$ separately.

Using the definition (3.8) of momenta, which implicitly contains the information about primary constraints, the variation satisfies

$$\delta H = \dot{q}^i \delta p_i + p_i \delta \dot{q}^i - \frac{\partial L}{\partial \dot{q}^i} \delta \dot{q}^i - \frac{\partial L}{\partial q^i} \delta q^i = \dot{q}^i \delta p_i - \frac{\partial L}{\partial q^i} \delta q^i \,. \tag{3.10}$$

Since the final expression, just as the general variation

$$\delta H = \frac{\partial H}{\partial q^i}\delta q^i + \frac{\partial H}{\partial p_i}\delta p_i \tag{3.11}$$

of a function on the momentum phase space, depends only on variations of q^i and p_j, while all other $\dot{q}^i$-variations cancel once the definition of momenta is used, H is guaranteed to be a well-defined function on the primary constraint surface. (As an example of a function that is not well defined on the primary constraint surface, consider $\dot{q}^i$. Its value is not determined for all i by just specifying a point (q^i, p_j) on the primary constraint surface. Accordingly, the variation along $\dot{q}^i$ can be expressed in terms of those along q^i and p_j only if $\dot{q}^i$ happens to be one of the velocities that can be expressed in terms of momenta, i.e. one of the velocities not primarily constrained.)

Returning to the Hamiltonian, the definition gives rise to a function $H(q^i, p_j)$ on the primary constraint surface whose variation, combining (3.10) and (3.11), satisfies the equation

$$\left(\frac{\partial H}{\partial q^i} + \frac{\partial L}{\partial q^i}\right)\delta q^i + \left(\frac{\partial H}{\partial p_i} - \dot{q}^i\right)\delta p_i = 0 \tag{3.12}$$

for any variation $(\delta q^i, \delta p_i)$ tangent to the primary constraint surface. Writing this equation as

$$\left(\frac{\partial H}{\partial q^i} + \frac{\partial L}{\partial q^i}\,,\, \frac{\partial H}{\partial p_j} - \dot{q}^j\right)\begin{pmatrix}\delta q^i \\ \delta p_j\end{pmatrix} = 0$$

shows that the vector

$$\delta := \left(\frac{\partial H}{\partial q^i} + \frac{\partial L}{\partial q^i}\,,\, \frac{\partial H}{\partial p_j} - \dot{q}^j\right)$$

must be normal to the constraint surface. Since the surface is represented as $\psi_s = 0$, $s = 1, \ldots, m$, a basis of its normal space is given by the gradients

$$v_s := \left(\frac{\partial \psi_s}{\partial q^i}\,,\, \frac{\partial \psi_s}{\partial p_j}\right)$$

of all the primary constraint functions. Thus, $\delta = \sum_s \lambda^s v_s$ for some coefficients λ^s (which might be functions on phase space), or

$$\frac{\partial H}{\partial q^i} + \frac{\partial L}{\partial q^i} = \lambda^s \frac{\partial \psi_s}{\partial q^i} \tag{3.13}$$

$$\frac{\partial H}{\partial p_i} - \dot{q}^i = \lambda^s \frac{\partial \psi_s}{\partial p_i}\,. \tag{3.14}$$

In this way, we have derived Hamiltonian equations of motion

$$\dot{q}^i = \frac{\partial H}{\partial p_i} - \lambda^s \frac{\partial \psi_s}{\partial p_i} \tag{3.15}$$

$$\dot{p}_i = \frac{\partial L}{\partial q^i} = -\frac{\partial H}{\partial q^i} + \lambda^s \frac{\partial \psi_s}{\partial q^i} \,. \tag{3.16}$$

Compared to the form of Hamiltonian equations of motion in unconstrained systems, there are new terms arising from the constraints. However, writing

$$\dot{q}^i = \frac{\partial (H - \lambda^s \psi_s)}{\partial p_i} + \frac{\partial \lambda^s}{\partial p_i} \psi_s \tag{3.17}$$

$$\dot{p}_i = -\frac{\partial (H - \lambda^s \psi_s)}{\partial q^i} - \frac{\partial \lambda^s}{\partial q_i} \psi_s \tag{3.18}$$

shows that, up to terms that vanish on the primary constraint surface, defined by $\psi_s = 0$, they can be seen as the usual Hamiltonian equations for the total Hamiltonian

$$H_{\mathrm{total}} = H - \lambda^s \psi_s \,. \tag{3.19}$$

In this case, one also writes

$$\dot{q}^i \approx \frac{\partial H_{\mathrm{total}}}{\partial p_i} \,, \quad \dot{p}_i \approx -\frac{\partial H_{\mathrm{total}}}{\partial q^i}$$

where the "weak equality" sign $\approx$ denotes an identity up to terms that vanish on the constraint surface.

While the value of the total Hamiltonian does not change by adding primary constraints and is independent of the λ^s, the evolution it generates depends on derivatives of the ψ_s. Unlike the ψ_s, the derivatives may not vanish, and evolution can thus depend on the (so far undetermined) λ^s. To see the role of the λ^s and how well-defined evolution can result, the mathematical theory of constraints, best described in terms of Poisson structures, is useful.

3.1.2.2 Poisson brackets

To discuss the general behavior of constrained systems, as well as those specific ones realized in general relativity, the concept of Poisson brackets and symplectic structures is the appropriate tool. As in classical mechanics, we have the general definition of Poisson brackets

$$\{f(q,p), g(q,p)\} := \sum_{i=1}^{n} \left(\frac{\partial f}{\partial q^i} \frac{\partial g}{\partial p_i} - \frac{\partial f}{\partial p_i} \frac{\partial g}{\partial q^i} \right) \tag{3.20}$$

for a system with finitely many degrees of freedom. (The generalization to infinite-dimensional cases will be done in Chapter 3.1.3.) As one can verify by direct calculations, Poisson brackets satisfy the following defining properties:

- linearity in both entries;
- antisymmetry when the two entries are commuted;

- the Leibniz rule $\{f, gh\} = \{f, g\}h + g\{f, h\}$;
- the Jacobi identity

$$\{\{f, g\}, h\} + \{\{g, h\}, f\} + \{\{h, f\}, g\} = 0\,. \tag{3.21}$$

With this definition, we can write the equations of motion in the compact form

$$\dot{q}^i \approx \{q^i, H_{\text{total}}\}\,, \qquad \dot{p}_i \approx \{p_i, H_{\text{total}}\} \tag{3.22}$$

or $\dot{F} \approx \{F, H_{\text{total}}\}$ for an arbitrary phase-space function $F(q^i, p_j)$. Whenever a change of phase-space variables is obtained in this way by taking Poisson brackets with a specific phase-space function H, H is said to *generate* the corresponding change. Viewing the variations $(\dot{q}^i, \dot{p}_j)$ as a vector field, the *Hamiltonian vector field* with components $\{\cdot, H\}$ is associated with any phase-space function H. In particular, (3.22) expresses the fact that the total Hamiltonian generates the dynamical flow of the phase-space variables in time.

Poisson tensor Poisson brackets present a geometrical notion of spaces that in several respects is quite similar to the notion of Riemannian geometry as it arises from a metric tensor; in others it is markedly different. If we express the Poisson bracket as

$$\{f, g\} = \mathcal{P}^{ij}(\partial_i f)(\partial_j g)\,, \tag{3.23}$$

which is always possible thanks to the linearity of Poisson brackets and the Leibniz rule, it becomes clear that the bracket structure is captured by a contravariant 2-tensor, or a bivector $\mathcal{P}^{ij}$, the *Poisson tensor*. Unlike a metric, this tensor is antisymmetric. Moreover, as a consequence of the Jacobi identity, it must satisfy

$$\epsilon_{ikl}\mathcal{P}^{ij}\partial_j\mathcal{P}^{kl} = 0\,. \tag{3.24}$$

Conversely, any antisymmetric 2-tensor satisfying (3.24) defines a Poisson bracket via (3.23). Since such tensors may depend non-trivially on the phase space coordinates, unlike the basic example of (3.20), a more general notion is obtained. The expression (3.20) represents the local form of a Poisson bracket as it can be achieved in suitably chosen phase-space coordinates q^i and p_j (so-called *Darboux coordinates*), and it is preserved by canonical transformations. But in general phase-space coordinates, the Poisson tensor need not be constant and may have components different from zero or ± 1.

Example 3.2 (Poisson tensor)

$$(\mathcal{P}^{ij}) := \begin{pmatrix} 0 & z & -y \\ -z & 0 & x \\ y & -x & 0 \end{pmatrix}$$

on $\mathbb{R}^3$*, i.e.* $\mathcal{P}^{ij} = \epsilon^{ijk}x_k$*, defines a Poisson tensor; see Exercise 3.1.*

The example of (3.20) leads to a Poisson tensor with an additional property: the matrix $\mathcal{P}^{ij}$ is invertible, with an inverse

$$\Omega_{ij} := (\mathcal{P}^{-1})_{ij}\,. \tag{3.25}$$

Invertibility is not a general requirement for Poisson tensors, as we will discuss in more detail below. But the non-degenerate case with an existing inverse occurs often and leads to several special properties. If the inverse Ω_{ij} of $\mathcal{P}^{ij}$ exists, providing an antisymmetric covariant 2-tensor, it is called a *symplectic form*.

As usual with metric tensors, we might be tempted to use the same symbol for an invertible Poisson tensor and its inverse 2-form, distinguished from each other only by the position of indices. However, this is normally not done for the following reason: a non-degenerate Poisson tensor defines a bijection $\mathcal{P}^\sharp \colon T^*M \to TM, \alpha_i \mapsto \mathcal{P}^{ij}\alpha_j$ from the co-tangent space to the tangent space of a manifold M, raising indices by contraction with $\mathcal{P}^{ij}$.[2] This is similar to the use of an inverse metric tensor, except that the Poisson tensor is antisymmetric. The inverse $\mathcal{P}^{\sharp-1}$ of this bijection defines a map $\mathbf{\Omega}^\flat = \mathcal{P}^{\sharp-1} \colon TM \to T^*M, v^i \mapsto (\mathcal{P})^{-1}_{ij} v^j = \Omega_{ij} v^j$ lowering indices. By tensor products, indices of tensors of arbitrary degree can be lowered and raised. In particular, we can lower the indices of $\mathcal{P}^{ij}$ using $\mathcal{P}^{\sharp-1}$:

$$(\mathcal{P}^{\sharp-1})_{ik}(\mathcal{P}^{\sharp-1})_{jl}\mathcal{P}^{kl} = \delta^l_i(\mathcal{P}^{\sharp-1})_{jl} = -(\mathcal{P}^{-1})_{ij}$$

which is not Ω_{ij} but its negative. The antisymmetry of tensors in Poisson geometry introduces an ambiguity which does not exist for Riemannian geometry. In Riemannian geometry, taking the inverse metric agrees with raising the indices of the metric by its inverse. In Poisson geometry, agreement is realized only up to a sign. For this reason, we do not follow the convention of using the same letter for the tensor and its inverse, as we would do for a metric, but rather, keep separate symbols $\mathcal{P}$ and Ω.

Symplectic form If we contract the Jacobi identity (3.24), written as

$$\mathcal{P}^{ij}\partial_j\mathcal{P}^{kl} + \mathcal{P}^{kj}\partial_j\mathcal{P}^{li} + \mathcal{P}^{lj}\partial_j\mathcal{P}^{ik} = 0\,,$$

with Ω_{mi} and Ω_{nk}, then use $\Omega_{ij}\partial_k\mathcal{P}^{jl} = -\mathcal{P}^{jl}\partial_k\Omega_{ij}$ due to the inverse relationship of the two tensors, we obtain[3]

$$\mathcal{P}^{lk}(\partial_m\Omega_{nk} - \partial_n\Omega_{mk} + \partial_k\Omega_{mn}) = \mathcal{P}^{lk}(\mathrm{d}\mathbf{\Omega})_{mnk} = 0\,.$$

Thus, the inverse of an invertible Poisson tensor is always a closed 2-form. It is also non-degenerate thanks to the invertibility, and any 2-form satisfying these properties is called a *symplectic form*. A Poisson structure $(M, \mathcal{P})$ on a manifold M with an invertible Poisson tensor is equivalent to a symplectic structure $(M, \mathbf{\Omega})$.

Hamiltonian vector fields Given a Poisson tensor, we associate the Hamiltonian vector field X_f to any function f on the Poisson manifold by

$$X_f := \mathcal{P}^\sharp \mathrm{d}f = -\mathcal{P}^{ij}(\partial_i f)\partial_j\,. \tag{3.26}$$

[2] This so-called “musical isomorphism” is sometimes denoted simply as $\sharp \colon T^*M \to TM$, with inverse $\flat \colon TM \to T^*M$. We prefer to indicate the object used to define the mapping, here the Poisson tensor, since general relativity employs also analogous but quite different mappings defined with a metric.

[3] Differential forms, whose tangent-space indices are suppressed, will be denoted here and throughout this book by bold-face letters.

(In Riemannian geometry, an analogous construction provides the normal vector to the surface given by $f = \text{const.}$) In terms of the Poisson bracket, one can write $X_f = \{\cdot, f\}$ as the action of the vector field on functions, to be inserted for the dot. The Poisson bracket itself can be written in terms of Hamiltonian vector fields and the symplectic form:

$$\mathbf{\Omega}(X_f, X_g) = \Omega_{ij}\mathcal{P}^{ik}\mathcal{P}^{jl}\partial_k f \partial_l g = -\mathcal{P}^{ki}\partial_k f \partial_i g = -\{f, g\}\,.$$

One can interpret the Hamiltonian vector field as the phase-space direction of change corresponding to the function f; for $f = p$, we have $X_p = \partial/\partial q$. When f is one of the canonical coordinates, its Hamiltonian vector field is along its canonical momentum. In this sense, the Hamiltonian vector field generalizes the notion of momentum to arbitrary phase-space functions. Integrating the Hamiltonian vector field X_f to a 1-parameter family of diffeomorphisms, we obtain the Hamiltonian flow $\Phi_t^{(f)}$ generated by f. The dynamical flow of a canonical system is generated by the Hamiltonian function on phase space.

Poisson and presymplectic geometry If the requirement of invertibility is dropped, two unequal siblings spring from the pair of Poisson and symplectic geometry. A non-invertible Poisson tensor is still said to provide a Poisson geometry, but it does not have an equivalent symplectic formulation. A closed 2-form not required to be invertible is called a *presymplectic form*, providing the manifold on which it is defined with presymplectic geometry.

A presymplectic form Ω_{ij} has a kernel $C \subset TM$ of vector fields v^i satisfying $\Omega_{ij}v^j = 0$. These vector fields define a flow on phase space, which can be factored out by identifying all points on orbits of the flow. The resulting factor space is symplectic: every vector field in the kernel of Ω_{ij} is factored out. In this way, a reduced symplectic geometry can be associated with any presymplectic geometry.

A degenerate Poisson tensor $\mathcal{P}^{ij}$ provides so-called *Casimir functions* C^I satisfying $\{C^I, f\} = 0$ with any other function f. Thus, the 1-forms $\mathrm{d}C^I$ are in the kernel of $\mathcal{P}^{ij}$, $\mathcal{P}^{ij}\partial_j C^I = 0$. In this situation, we naturally define a distribution $T \subset TM$ of subspaces such that $\mathrm{d}C^I(v) = v^i\partial_i C^I = 0$ for any $v^i \in T$. Any v^i in this distribution is annihilated by the kernel of $\mathcal{P}^{ij}$ and is thus in the image, expressed as $v^i = \mathcal{P}^{ij}\alpha_j$ for some co-vector α_i; conversely, any $\mathcal{P}^{ij}\alpha_j$ satisfies $\mathcal{P}^{ij}\alpha_j\partial_i C^I = 0$ and is thus in T. Thus, $T = \mathrm{Im}\mathcal{P}^\sharp$. Thanks to the Jacobi identity (3.24), this distribution is integrable: for two vectors $v^i = \mathcal{P}^{ij}\alpha_j$ and $w^i = \mathcal{P}^{ij}\beta_j$ in the distribution, the Lie bracket $[v, w]^i$ is also in the distribution. Indeed, a quick calculation using the Jacobi identity shows that

$$\begin{aligned}
[v, w]^i &= \mathcal{P}^{jk}\alpha_k\partial_j(\mathcal{P}^{il}\beta_l) - \mathcal{P}^{jk}\beta_k\partial_j(\mathcal{P}^{il}\alpha_l) \\
&= \mathcal{P}^{il}\left(\mathcal{P}^{jk}(\alpha_k\partial_j\beta_l - \beta_k\partial_j\alpha_l) + \alpha_k\beta_j\partial_l\mathcal{P}^{jk}\right) =: \mathcal{P}^{il}\gamma_l
\end{aligned}$$

is an element of the distribution. Thanks to the integrability, a Poisson manifold is foliated[4] into submanifolds spanned by the distribution, with tangent spaces annihilated by all Casimir functions. Integral surfaces of the distribution, leaves of this foliation, are called

[4] In general, the foliation may be singular: leaves are not guaranteed to have the same dimension.

symplectic leaves. The leaves are indeed symplectic: the Poisson tensor restricted to the space co-tangent to the leaves is invertible.

Example 3.3 (Symplectic leaves)
Define $\mathcal{P}^{ij} = \epsilon^{ijk} x_k$ on $\mathbb{R}^3$, as in Example 3.2. The conditions $\mathcal{P}^{ij}\partial_j C = 0$ for a Casimir function C impose the equations $\epsilon^{ijk}\partial_k C = 0$, which is solved by any function depending on x, y and z only via $x^2 + y^2 + z^2$. We can thus choose our Casimir function as $C(x, y, z) = x^2 + y^2 + z^2$. Surfaces of constant $C = r^2$ are spheres of radius r, on which it is useful to choose polar coordinates φ and ϑ. With the relationship between polar and Cartesian coordinates, one finds that $\mathcal{P}(\mathrm{d}\varphi, \sin\vartheta \mathrm{d}\vartheta) = \sin\vartheta \mathcal{P}^{ij}\partial_i\varphi\partial_j\vartheta = r^{-1}$ is non-vanishing for $r \neq 0$, and thus the Poisson tensor is non-degenerate on the co-tangent space of a sphere. Spheres are symplectic leaves of the Poisson manifold considered here.

Constraints on symplectic manifolds If we constrain a symplectic manifold M with symplectic form Ω_{ij} and Poisson tensor $\mathcal{P}^{ij}$ to a subset $\mathcal{C}$ defined by the vanishing of constraint functions C^I, $\mathcal{C}$: $C^I = 0$, there are different possibilities for symplectic properties of the subset. These properties are mainly determined by the Hamiltonian vector fields of the constraints:[5] We call a constraint C^I *first class* with respect to all constraints if its Hamiltonian vector field is everywhere tangent to the constraint surface $\mathcal{C}$. We call it *second class* if its Hamiltonian vector field is nowhere tangent to the constraint surface. By the definition of Hamiltonian vector fields, this is equivalent to saying that $\{C^I, C^J\}$ vanishes on the constraint surface for all C^J if C^I is first class; it vanishes nowhere on the constraint surface if C^I is second class. (No condition is posed for the behavior of these Poisson brackets off the constraint surface.) The surface is called a *first-class constraint surface* if all constraints defining it are first class, and a *second-class constraint surface* if all constraints are second class. (The terminology of first- and second-class constraints was introduced by Anderson and Bergmann (1951), developed further by Dirac (1958a) and especially for gravity by Dirac (1958b), then summarized by Dirac (1969).)

We can equip the constraint surface with a presymplectic form $\bar{\boldsymbol{\Omega}}$ by pulling back the symplectic form to it. If $\iota\colon \mathcal{C} \to M$ is the embedding of the constraint surface in M, we write $\bar{\boldsymbol{\Omega}} = \iota^*\boldsymbol{\Omega}$. In components, if $\iota\colon\ y^\alpha \mapsto x^i$ describes the embedding in local coordinates, $\bar{\Omega}_{\alpha\beta} = \Omega_{ij}(\partial x^i/\partial y^\alpha)(\partial x^j/\partial y^\beta)$. (We cannot directly equip it with a Poisson structure, since the Poisson tensor, being contravariant, cannot be pulled back.) If one of the constraints is first class, say C, the presymplectic form is degenerate: its Hamiltonian vector field X_C^i is tangent to the constraint surface and thus defines a vector field X^α on it by simple restriction. For this, we have $\bar{\Omega}_{\alpha\beta}X^\beta = \iota^*(\Omega_{ij}X_C^j) = \iota^*(\partial_i C) = 0$ using the definition of Hamiltonian vector fields.

Only if all constraints are second class does a symplectic structure result on the constraint surface. In this case, it can directly be used as the phase space of the reduced system where

[5] For all constraints considered here, we assume the Hamiltonian vector field to be non-vanishing in a neighborhood of the constraint surface. Such constraints, called regular, provide good local coordinates transversal to the constraint surface.

the constraints are solved. If there are first-class constraints, their Hamiltonian flow must be factored out to obtain the *reduced phase space* as the factor space of the presymplectic constraint surface by the Hamiltonian flow. In physical terms, this flow is the gauge flow generated by the constraints as discussed in more detail below.

Dirac bracket If one does not explicitly solve all the constraints, which can often be complicated, one may work with the constrained system implicitly by using the constraints without solving them. In this case, if Hamiltonian flows of phase-space functions are used, most importantly the dynamical flow generated by the Hamiltonian, one must be careful that they do not leave the constraint surface. This would be guaranteed if all constraints are solved explicitly — every flow would automatically be restricted to the reduced phase space — but not if some of them cannot be solved. If all constraints are first class, their Hamiltonian flow is tangent to the constraint surface; they do not cause a problem in this context. But second-class constraints generate a flow transversal to the constraint surface. If second-class constraints are left unsolved, they may contribute terms to a Hamiltonian making the flow move off the constraint surface.

Rather than solving all second-class constraints, one may modify the Poisson brackets so as to ensure that the Hamiltonian flow generated by the old constraints with respect to the new Poisson structure is tangent to the constraint surface. This is possible by introducing the *Dirac bracket*

$$\{f,g\}_{\mathrm{D}} := \{f,g\} - \sum_{I,J}\{f,C^I\}(\{C^I,C^J\})^{-1}\{C^J,g\}\,. \tag{3.27}$$

Here, the double sum is taken for all second-class constraints, for which the matrix inverse of $\{C^I, C^J\}$ is guaranteed to exist. With this new Poisson bracket, the flow generated by the second-class constraints does indeed not leave the constraint surface; it even vanishes:

$$\{f,C^K\}_{\mathrm{D}} = \{f,C^K\} - \sum_{I,J}\{f,C^I\}(\{C^I,C^J\})^{-1}\{C^J,C^K\} = 0\,.$$

The new Poisson tensor defined by the Dirac bracket is degenerate, with the second-class constraints as its Casimir functions. Moreover, for any phase-space function f whose Hamiltonian vector field is already tangent to the second-class constraint surface with the original Poisson structure, i.e. $X_f C^I = 0$ for all second-class constraints C^I, we have $\{f, C^I\} = \mathcal{P}^{ij}\partial_i f \partial_j C^I = -X_f^j \partial_j C^I = -X_f C^I = 0$, and thus $\{f,g\}_{\mathrm{D}} = \{f,g\}$ with any other phase-space function g. In summary, the Dirac bracket leaves the Poisson structure for functions generating a flow tangent to the constraint surface unchanged, while removing any flow off the constraint surface generated by functions of the second-class constraints. Even if not all second-class constraints can be solved, the Dirac bracket, which is often easier to compute, makes sure that they do not lead to spurious flows off the constraint surface. (Further mathematical properties of the Dirac bracket are discussed in the Appendix.)

Using the Dirac bracket, only the flow of first-class constraints need be considered. It cannot be removed in a similar way, since the matric $\{C^I, C^J\}$ is not invertible if first-class constraints are present. To obtain a phase space with symplectic structure, the flows generated by first-class constraints must be factored out, as already stated. We will soon see why this factoring-out is necessary from a physical perspective.

3.1.2.3 Constraint algebras

With information about Poisson geometry and the mathematical theory of constraint surfaces in symplectic spaces, we now return to the physical analysis of constrained systems and their dynamics. We had arrived at the total Hamiltonian, $H_{\rm total}$ in (3.19), as the sum of the Hamiltonian and all primary constraints ψ_s multiplied with parameters λ^s. The λ^s play the role of Lagrange multipliers which add the primary constraints to the original Hamiltonian, making sure that these constraints are implemented. They may or may not be determined by solving the equations of motion or other equations, depending on the form of the constraints. Here, the mathematical classification of constraints becomes relevant.

Secondary constraints In addition to the primary constraints, there are further equations to be satisfied by initial values: Consistency conditions are required because the time derivatives of the primary constraints, like the constraints themselves, must vanish at all times,

$$0 = \dot{\psi}_s \approx \{\psi_s, H\} - \lambda^t\{\psi_s, \psi_t\} =: \{\psi_s, H\} - \lambda^t C_{st}\,. \tag{3.28}$$

(In the weak equality we are again free to move the λ^t past the derivatives contained in the Poisson bracket because only terms constrained to be zero result in this way.)

The (phase-space dependent) matrix $C_{st} := \{\psi_s, \psi_t\}$ introduced here determines the structure of the constrained system. If $\det(C_{st}) \neq 0$, no further constraints result and we can fulfill the consistency conditions (3.28) by solving them for all λ^t. In this way, the total Hamiltonian and the flow it generates will be completely specified. If $\det(C_{st}) = 0$, however, not all λ^t can be determined to solve the consistency conditions completely. In this case, (3.28) implies *secondary constraints* which follow from the equations of motion, rather than from the basic definition of momenta as the primary constraints. Secondary constraints in general take the form $Z_r^s\{\psi_s, H\} = 0$ for any zero-eigenvector Z_r^s of the primary constraint matrix: $Z_r^s C_{st} = 0$. If a non-trivial zero-eigenvactor Z_r^s exists, $\psi_s = 0$ is preserved in time only if $Z_r^s\{\psi_s, H\} = 0$.

Also, the secondary constraints imply consistency conditions of the form (3.28), which may generate further constraints. This process must be continued until all consistency conditions for all the constraints are satisfied, making sure that constraints can hold at all times. Once this is realized, one has determined the complete constrained system. At this stage, the distinction between primary and other constraints becomes irrelevant, and so we call all constraints derived in this way C_s, including the primary constraints ψ_s. (For simplicity, we reuse the label s, even though it may now run through a larger range.) If $\det(\{C_s, C_t\})$ does not vanish on the constraint surface, the constraints are *second class*; more generally, the number of second-class constraints is given by $\mathrm{rank}(\{C_s, C_t\})$. If all constraints are second class, we can solve for all the multipliers λ^t and the dynamics is determined uniquely.

Example 3.4 (Classical mechanics on a surface)
Consider a system of two coordinates q^1 and q^2 with momenta p_1, p_2 whose dynamics is given by the Hamiltonian

$$H = \frac{1}{2}(p_1^2 + p_2^2) + V(q^1, q^2) + \lambda f(q^1, q^2)\,. \tag{3.29}$$

The two functions V and f, respectively, play the role of the potential and a constraint forcing the motion to be on the surface $f(q^1, q^2) = 0$. Since no momentum p_λ appears in the Hamiltonian, we have one primary constraint $C_1 := p_\lambda = 0$. It implies the secondary constraint $C_2 := -\{p_\lambda, H\} = f(q^1, q^2) = 0$, which implies a tertiary constraint

$$C_3 := \{f, H\} = \frac{\partial f}{\partial q^1} p_1 + \frac{\partial f}{\partial q^2} p_2 = 0\,,$$

which in turn implies a quartary constraint

$$\begin{aligned} C_4 := \{C_3, H\} = {} & \frac{\partial^2 f}{\partial (q^1)^2} p_1^2 + 2\frac{\partial^2 f}{\partial q^1 \partial q^2} p_1 p_2 + \frac{\partial^2 f}{\partial (q^2)^2} p_2^2 \\ & -\frac{\partial f}{\partial q^1}\left(\frac{\partial V}{\partial q^1} + \lambda \frac{\partial f}{\partial q^1}\right) - \frac{\partial f}{\partial q^2}\left(\frac{\partial V}{\partial q^2} + \lambda \frac{\partial f}{\partial q^2}\right) = 0\,. \end{aligned}$$

The secondary constraint requires all positions to be on the surface $f = 0$, and the tertiary one ensures that the momentum vector is tangent to this surface. (The constraint can be written as $C_3 = \vec{N} \cdot \vec{p} = 0$ with the (unnormalized) vector $\vec{N} = \nabla f$ normal to the surface.) The last constraint then is equivalent to a balance equation between the normal component $\vec{N} \cdot \vec{F}$ of the force $\vec{F} = -\nabla V - \lambda \nabla f$ obtained as the sum of the external force due to the potential and a force exerted by the surface, and a centrifugal force given in terms of second-order derivatives of f and momenta. The form of the centrifugal force obtained here for a general surface is not easy to recognize, but if we specialize the example to a circle of radius R, $f(q^1, q^2) = (q^1)^2 + (q^2)^2 - R^2$, the centrifugal force reduces to $R^{-1}\vec{p}^2$, the usual form for a particle of unit mass. (We must divide the expression in C_4 by the norm of $\vec{N}$, in this case $2R$, in order to read off the properly normalized mass. We will see in Chapter 3.2.2.2 that first-order derivatives of the normal vector as encountered here provide a measure for the extrinsic curvature of the surface embedded in space.) The surface force $-\lambda \nabla f$ is determined once we solve $C_4 = 0$ for λ, which is always possible, since $\vec{N}^2 = (\partial f/\partial q^1)^2 + (\partial f/\partial q^2)^2 \neq 0$ for a regular surface. At this stage, we see that at least some of the constraints must be second class, since we were able to solve uniquely for the multiplier.

In fact, a derivation of the constraint algebra shows that all constraints in this example are second class. We easily compute $\{C_1, C_4\} = \vec{N}^2 = \{C_2, C_3\}$, and clearly C_1 has vanishing Poisson brackets with both C_2 and C_3. The Poisson brackets $\{C_2, C_4\}$ and $\{C_3, C_4\}$ are more lengthy, but, irrespective of their value, the constraint matrix $\{C_s, C_t\}$ has determinant $\vec{N}^4 \neq 0$, proving that the constraints are second class.

If the constraints satisfy $\{C_s, C_t\} \approx 0$ on the constraint surface $C_t = 0$, they are *first-class constraints*. (In the presence of a Hamiltonian generating evolution, one usually includes $\{C_s, H\} = 0$ as a condition for first-class constraints combined with a first-class Hamiltonian. One may interpret this as including the condition of energy conservation with the constraints C_s.) Not all the Lagrange multipliers can be solved for, and the dynamics is not unique. Gauge transformations arise.

Gauge transformations As seen in the general analysis of Poisson geometry, first-class constraints play a special role because their constraint surface is not symplectic: they generate a Hamiltonian flow tangent to the constraint surface, along which the presymplectic form, obtained by pulling the symplectic form back to the constraint surface, is degenerate. In the physical interpretation, this means that first-class constraints not only restrict initial values but also generate *gauge transformations*. The infinitesimal mapping

$$F(q,p) \mapsto F(q,p) + \delta^{(s)}_{\epsilon} F(q,p) := F(q,p) + \{F, \epsilon C_s\} \tag{3.30}$$

of a phase-space function F, defined for any first-class constraint C_s, maps solutions to the constraints and equations of motion into other solutions: under this mapping, $\delta^{(s)}_{\epsilon} C_t \approx 0 \approx \delta^{(s)}_{\epsilon} H$.

Interpreting these transformations as gauge means that we do not consider solutions mapped to each other by the Hamiltonian flow of first-class constraints as physically distinct. We are not merely applying a symmetry transformation to map a known solution into a new one, like rotating an orbit in a spherically symmetric potential which gives rise to a new allowed solution. Gauge transformations map different mathematical solutions into each other, but they are interpreted merely as different representations of the same physical solution.

This stronger interpretation is required when we consider the dynamical flow generated by the total Hamiltonian $H_{\mathrm{total}} = H_1 - \lambda^s \psi_s$, which explicitly contains the primary constraints ψ_s, but also implicitly secondary or other ones in H_1. (We now use the symbol H_1 to indicate that second-class constraints should already have been solved at this stage, fixing their multipliers. Alternatively, we could use the Dirac bracket in the following argument.) Every phase-space function f changes in time by $\dot{f} = \{f, H_{\mathrm{total}}\}$. In general, the change in time of a function f then depends on the undetermined parameters λ^s corresponding to the first-class constraints unless $\{f, C_s\} = 0$ for all first-class constraints. If f has a vanishing Poisson bracket with all first-class constraints, it is called a complete (or *Dirac*) *observable*. For phase-space functions with non-vanishing Poisson brackets with the first class constraints, however, a dependence of their change in time on some of the λ^s cannot be avoided. To have a well-defined theory with unambiguous physical predictions, the only viable conclusion is that the infinitesimal flow $\{f, C_s\}$ only changes the mathematical representation but does not change the physics of the observable information. The change of f is just a gauge transformation without affecting the physical state.

Example 3.5 (Gauge in isotropic models)
In Chapter 2.1 we saw an example of a system with two constraints, the primary constraint $p_N = 0$ *and the secondary constraint* $C = 0$. *Since* C *is independent of* N, *we immediately have* $\{p_N, C\} = 0$ *such that the system of constraints is first class. The total Hamiltonian is* $NC - \lambda p_N$ *where* N *arises automatically as a multiplier of the secondary constraint* C.

The gauge transformations generated by the constraints correspond to time reparameterizations, as one can see explicitly: we have

$$\delta_\epsilon a = \{a, \epsilon NC\} = -\frac{4\pi G}{3}\epsilon \frac{p_a}{a} N = \epsilon \dot{a}$$

and $\delta_\lambda N = \{N, \lambda p_N\} = \lambda$, $\delta_\epsilon p_a = \{p_a, \epsilon NC\} = \epsilon \dot{p}_a$ *and finally* $\delta_\epsilon p_N = C \approx 0$. *The gauge transformation* $a \mapsto a + \delta_\epsilon a = a + \epsilon \dot{a} = a(t+\epsilon) + O(\epsilon^2)$ *thus agrees to linear order in* ϵ *with a change of the time coordinate while* N *can be changed arbitrarily. One may view this as evolution along the time parameter, or as a reparameterization of the old time parameter* t *to a new one making coordinate time gauge-dependent.*

In totally constrained systems such as the preceding example, constraints play several roles: they constrain allowed field values to reside on the constraint surface, they generate gauge transformations as identifications of physically equivalent field configurations on the constraint surface and they provide the total Hamiltonian. The last property means that a particular combination of the constraints, such as $H = NC$ in the example with a single constraint, generates Hamiltonian equations of motion in a time coordinate. Which combination we choose for the total Hamiltonian, i.e. which multiplier function N we use in the example, will have an influence on the form of equations of motion. For first-class constraints, this function remains undetermined by the constraints. Any choice is allowed, and there is no unique, absolute notion of evolution — a further reflection of the gauge structure of the theory. Any particular choice for the total Hamiltonian will result in equations of motion written in a specific gauge. But since the theory is invariant under gauge transformations generated by the constraints, the choice of a total Hamiltonian does not matter, and all sets of equations of motion obtained for different gauges are equivalent. We will see in this chapter that general relativity, also in the absence of space-time symmetries, has first-class constraints and a fully constrained Hamiltonian. We will also encounter formulations that have additional second-class constraints.

3.1.3 Field theories

Lagrangian or Hamiltonian formulations of field theories follow the same lines as discussed so far for finite-dimensional systems. One may view such theories as having independent variables $\varphi(x)$ labelled not by a discrete parameter i taking a finite number of values, but by a continuous parameter x. Most constructions still go through with this understanding, but it hides several subtleties. Especially, derivatives by the fields, as they are required for variational equations or Poisson brackets, require mathematical care. For instance, instead of having the relationship $\partial q^i/\partial q^j = \delta^i_j$ for independent variables, a basic *functional* *derivative* as it appears for field theories is $\delta\varphi(x)/\delta\varphi(y) = \delta(x, y)$ where the new type of derivative is denoted by the variational symbol δ and $\delta(x, y)$ is Dirac's delta-distribution.

If we have a functional $S[\varphi]$ defined for fields on an n-dimensional manifold (we will be applying functional derivatives to fields on space-time and on space) whose *variation* under small changes of φ, the part of $S[\varphi + \delta\varphi] - S[\varphi]$ linear in $\delta\varphi$, has the form

$$\delta S = \int \mathrm{d}^n x A[\varphi(x)]\delta\varphi(x) \tag{3.31}$$

then S is called *functionally differentiable by* φ with the *functional derivative* $\delta S/\delta\varphi(x) = A[\varphi(x)]$ of S by φ.

Formally, all rules for differentiation, such as linearity and the product and chain rules, can be extended to functional derivatives. Functional derivatives of spatial derivatives of a function can be computed by using integration-by-parts formulas, for instance

$$\frac{\delta}{\delta\varphi(x)} \int \mathrm{d}^3 y N^a(y) p_\varphi(y) \partial_a \varphi(y) = -\partial_a (N^a(x) p_\varphi(x))$$

as a specific example (the diffeomorphism constraint) which we will later see for a scalar field theory with scalar field φ and momentum p_φ independent of φ (and certain space-time metric components N^a). Here, we have ignored boundary terms by appealing to suitable boundary conditions to be satisfied for the fields. Such boundary conditions determine whether a functional of the fields is functionally differentiable: would a boundary term remain in the variation, δS would not be of the required form (3.31). Applying rules such as integration by parts is sufficient to set up the full framework as it exists for finite-dimensional systems, but care is still required due to the occurrence of delta-distributions.

To avoid the explicit use of distributions and instead work with well-defined algebraic relationships, one can "smear" fields or functionals of them. Instead of working with local combinations of fields, one then considers integrals such as $\varphi[\mu] := \int \mathrm{d}^n x \sqrt{|\det g|} \mu(x) \varphi(x)$ on a curved manifold with metric g_{ab} for arbitrary functions $\mu(x)$. No information is lost, since $\varphi(x)$ can be reconstructed if $\varphi[\mu]$ is known for all (smooth, say) $\mu(x)$. Derivatives by fields

$$\frac{\delta\varphi[\mu]}{\delta\varphi(x)} = \sqrt{|\det g|}\mu(x) \tag{3.32}$$

are free of delta-distributions. We may view smearing functions as a different set of labellings replacing the continuous spatial points x where field values are evaluated.

Similarly, Poisson brackets of smeared functionals become well-defined algebraic relationships. For canonical fields, Poisson brackets are defined using functional derivatives in

$$\{f, g\} = \int \mathrm{d}^n x \left(\frac{\delta f}{\delta\varphi(x)} \frac{\delta g}{\delta p_\varphi(x)} - \frac{\delta f}{\delta p_\varphi(x)} \frac{\delta g}{\delta\varphi(x)} \right) \tag{3.33}$$

for a scalar field with its momentum, and analogously for tensorial fields. Instead of

$$\{\varphi(x), p_\varphi(y)\} = \delta(x, y)$$

for the fields, we have

$$\{\varphi[\mu], p_\varphi(y)\} = \sqrt{|\det g|}\mu(y)$$

(a function, not a distribution) if at least one of them is smeared.

In the context of smearing, the *density weight* of fields is important, since it determines the metric factors $\sqrt{|\det g|}$ required for well-defined integration: the measure $\mathrm{d}^n x \sqrt{|\det g|}$ is invariant under changes of coordinates thanks to the transformation properties $\sqrt{|\det g|}' = |\det(\partial x^\mu/\partial x'^{\nu'})|\sqrt{|\det g|}$ that follow from the transformation $g_{\mu'\nu'} = (\partial x^\mu/\partial x'^{\mu'})(\partial x^\nu/\partial x'^{\nu'})g_{\mu\nu}$ of the metric tensor under a change $x^\mu \mapsto x'^{\mu'}$ of coordinates. Any function $\tilde{X}$ transforming by $\tilde{X}' = |\det(\partial x^\mu/\partial x'^{\nu'})|\tilde{X}$, even if it is not derived from the determinant of a metric, is called a *scalar density* (of weight one), and any tensor that transforms with an extra factor of $|\det(\partial x^\mu/\partial x'^{\nu'})|$ is called a *tensor density* (of weight one). Scalar densities of weight one can directly be integrated without an extra metric factor. More information on densities can be found in the Appendix.

When we smeared the scalar field, we explicitly included the metric factor in addition to the smearing function. The momentum of the scalar, however, must behave differently, since it appears together with the scalar in the symplectic term $\int \mathrm{d}^3 x \dot{\varphi} p_\varphi$ of the Lagrangian, and no other field, not even the metric, is allowed in this term, for otherwise p_φ would not be canonically conjugate to φ. The momentum p_φ of a scalar must then be a scalar density such that $\int \mathrm{d}^3 x p_\varphi$ is already coordinate invariant. Indeed, if we compute the momentum from the action (see Eq. (2.16)), we obtain $p_\varphi = \delta S/\delta\dot{\varphi} = -g^{00}\sqrt{|\det g|}\dot{\varphi}$ as a scalar density. In any canonical theory, tensorial configuration variables fully contracted with their momenta must be densities of weight one.

Example 3.6 (Electromagnetism on Minkowski space)
The action of Maxwell theory on Minkowski space-time with metric $\eta_{\mu\nu}$ is

$$S_{\mathrm{Maxwell}}[A_\mu] = -\frac{1}{4}\int \mathrm{d}^4 x\, F_{\mu\nu}(A)F_{\rho\sigma}(A)\eta^{\mu\rho}\eta^{\nu\sigma} \tag{3.34}$$

where $F_{\mu\nu}(A) = \partial_\mu A_\nu - \partial_\nu A_\mu$. First, we have

$$\frac{\delta S_{\mathrm{Maxwell}}}{\delta \partial_\mu A_\nu} = -(\partial_\rho A_\sigma - \partial_\sigma A_\rho)\eta^{\rho[\mu}\eta^{\nu]\sigma}$$

and thus, specializing to $\mu = 0$,

$$\frac{\delta S_{\mathrm{Maxwell}}}{\delta \dot{A}_\nu} = -(\dot{A}_\sigma - \partial_\sigma A_0)\eta^{\nu\sigma}\,. \tag{3.35}$$

If $\nu = 0$, the right-hand side vanishes identically; the action does not depend on the time derivative of A_0, making its momentum $\delta S_{\mathrm{Maxwell}}/\delta\dot{A}_0$ a primary constraint. The secondary constraint that it implies is the Gauss law

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\delta S_{\mathrm{Maxwell}}}{\delta \dot{A}_0} = \frac{\delta S_{\mathrm{Maxwell}}}{\delta A_0} = -\partial_a(\partial_b A_0 - \dot{A}_b)\delta^{ab} = \partial_a E^a \approx 0$$

with the electric field $E^a = \delta^{ab}(\dot{A}_b - \partial_b A_0)$ *whose components, as seen from (3.35), provide the momenta of* A_a*:* $\{A_a(x), E^b(y)\} = \delta_a^b \delta(x, y)$*. (Only spatial indices contribute at this stage. Note that the fixed background metric prevents all density weights from being seen.)*

The Gauss constraint does not generate further constraints, and so the system is complete: on our phase space we can directly eliminate the simple primary constraint function, leaving as canonical variables A_a *with momenta* E^b*, subject to the Gauss constraint* $\partial_a E^a = 0$*. The Gauss constraint does not depend on* A_a*, and is thus first class. If we smear it to* $G[\Lambda] = \int \mathrm{d}^3x \Lambda(x)\partial_a E^a$*, we have the Abelian algebra* $\{G[\Lambda_1], G[\Lambda_2]\} = 0$*. The smeared constraint generates gauge transformations leaving* E^a *fixed and changing* A_a *by the gradient* $\{A_a, G[\Lambda]\} = -\partial_a \Lambda$ *of* Λ*.*

If we introduce a gauge-fixing condition, for instance the Coulomb gauge $C := \delta^{ab}\partial_a A_b = 0$ *smeared to* $C[K] = \int \mathrm{d}^3x K(x)C(x)$*, the system of constraints ceases to be first class. We have*

$$\{G[\Lambda], C[K]\} = \int \mathrm{d}^3x \mathrm{d}^3y (\partial_a \Lambda)(x)(-\delta^{ab}\partial_b K(y))\delta(x, y) = \int \mathrm{d}^3x K \delta^{ab}\partial_a \partial_b \Lambda$$

which clearly does not vanish everywhere on the constraint surface. The system provided by the Gauss constraint together with the Coulomb gauge condition is second class for all smearing functions satisfying $\delta^{ab}\partial_a\partial_b\Lambda \neq 0$*. For harmonic functions satisfying* $\delta^{ab}\partial_a\partial_b\Lambda = 0$*, the gauge fixing is not complete and does not provide a second-class system of constraints. This is the well-known fact that the Coulomb gauge fixes the gauge freedom of changing* A_a *to* $A_a - \partial_a \Lambda$ *only up to gauge transformations generated by harmonic functions: if* $\delta^{ab}\partial_a A_b = 0$ *satisfies the gauge, also* $\delta^{ab}\partial_a(A_b - \partial_b\Lambda) = 0$ *satisfies it if and only if* $\delta^{ab}\partial_a\partial_b\Lambda = 0$*.*

In general, completely fixing the gauge requires a restriction to a subspace intersecting each gauge orbit exactly once. If this can be achieved, which is often difficult to do in explicit form, no gauge freedom is left. Since all gauge flows are then transversal to the gauge-fixing surface, the set of constraints plus gauge-fixing conditions is second class by definition.

Example 3.7 (Einstein–Hilbert action)
For gravity, the action to start with is the Einstein–Hilbert action

$$S_{\mathrm{EH}}[g] = \int \mathrm{d}^4x \left(\frac{\sqrt{-\det g}R}{16\pi G} + \mathcal{L}_{\mathrm{matter}} \right) \tag{3.36}$$

whose stationary points are given by solutions to Einstein's equation. The proof is lengthy but makes use of several identities which will also appear in the canonical formulation. Before continuing with the machinery of Hamiltonian formulations, we will thus first demonstrate explicitly that (3.36) is indeed the correct starting point to derive momenta and equations of motion.

We compute the variation $\delta(\sqrt{-\det g}R_{ab}g^{ab})$ *by a background-field expansion, setting* $g_{ab} = \tilde{g}_{ab} + \delta g_{ab}$ *for an arbitrary* $\tilde{g}_{ab}$ *and expanding to linear order in* δg_{ab}*. Functional* *derivatives can then be read off using (3.31). We will show that $\tilde{g}_{ab}$ satisfies Einstein's equation if and only if $\delta S_{\rm EH} = 0$.*

1. *Connection components. We call ∇_a the derivative operator compatible with g_{ab}, and $\tilde{\nabla}_a$ the operator compatible with $\tilde{g}_{ab}$. All indices will be lowered or raised with g_{ab} and its inverse. This gives*

$$0 = \nabla_a g_{bc} = \tilde{\nabla}_a g_{bc} - C^d_{ab} g_{dc} - C^d_{ac} g_{bd}$$

for a partially symmetric tensor $C^a_{bc} = C^a_{(bc)}$. Thus, $C_{cab} + C_{bac} = \tilde{\nabla}_a g_{bc}$ for $C_{abc} := g_{ad} C^d_{bc}$ and, using the symmetry of C^a_{bc},

$$\begin{aligned} C^c_{ab} &= g^{cd} C_{dab} = \frac{1}{2} g^{cd} \left((C_{dab} + C_{bad}) + (C_{dba} + C_{abd}) - (C_{bda} + C_{adb}) \right) \\ &= \frac{1}{2} g^{cd} \left(\tilde{\nabla}_a g_{bd} + \tilde{\nabla}_b g_{ad} - \tilde{\nabla}_d g_{ab} \right) \\ &= \frac{1}{2} (\tilde{g}^{cd} + \delta g^{cd}) \left(\tilde{\nabla}_a \delta g_{bd} + \tilde{\nabla}_b \delta g_{ad} - \tilde{\nabla}_d \delta g_{ab} \right) . \end{aligned}$$

Due to $C^c_{ab}|_{\delta g_{ab}=0} = 0$, this is at least of linear order in δg_{ab}. Its linear term is

$$\delta C^c_{ab} = \frac{1}{2} \tilde{g}^{cd} \left(\tilde{\nabla}_a \delta g_{bd} + \tilde{\nabla}_b \delta g_{ad} - \tilde{\nabla}_d \delta g_{ab} \right) . \tag{3.37}$$

2. *Curvature. Using $\nabla_a \nabla_b \omega_c - \nabla_b \nabla_a \omega_c = R_{abc}{}^d \omega_d$ for any 1-form ω_a and*

$$\nabla_a \nabla_b \omega_c = \tilde{\nabla}_a \tilde{\nabla}_b \omega_c - \tilde{\nabla}_a (C^d_{bc} \omega_d) - C^e_{ab} (\tilde{\nabla}_e \omega_c - C^d_{ec} \omega_d) - C^e_{ac} (\tilde{\nabla}_b \omega_e - C^d_{be} \omega_d)$$

we have

$$(\nabla_a \nabla_b - \nabla_b \nabla_a) \omega_c = (\tilde{\nabla}_a \tilde{\nabla}_b - \tilde{\nabla}_b \tilde{\nabla}_a) \omega_c - (\tilde{\nabla}_a C^d_{bc} - \tilde{\nabla}_b C^d_{ac}) \omega_d + (C^e_{ca} C^d_{be} - C^e_{cb} C^d_{ae}) \omega_d$$

and thus

$$R_{abc}{}^d = \tilde{R}_{abc}{}^d - 2 \tilde{\nabla}_{[a} C^d_{b]c} + 2 C^e_{c[a} C^d_{b]e} .$$

For the Ricci tensor, this implies that

$$R_{ac} = \tilde{R}_{ac} - 2 \tilde{\nabla}_{[a} C^b_{b]c} + 2 C^e_{c[a} C^b_{b]e}$$

which has as linear variation

$$\delta R_{ac} = - \tilde{\nabla}_a \delta C^b_{bc} + \tilde{\nabla}_b \delta C^b_{ac} . \tag{3.38}$$

In particular, $\tilde{g}^{ac} \delta R_{ac} =: \nabla_a v^a$ is a total divergence.

3. *Determinant. We make use of the identity*

$$\varepsilon^{abcd} g_{ac} g_{bf} g_{cg} g_{dh} = \bar{\varepsilon}_{efgh} \det g \tag{3.39}$$

where ε^{abcd} and $\bar{\varepsilon}_{abcd}$ both denote the totally antisymmetric tensor taking values zero or ± 1. (We use the bar to indicate that $\bar{\varepsilon}_{abcd}$ is not obtained by lowering indices of ε^{abcd}. In fact, as tensor fields they are quite different, ε^{abcd} having density weight one and $\bar{\varepsilon}_{abcd}$ density weight minus one;

see the Appendix.) Thus,

$$\begin{aligned}\delta \det g &= \frac{1}{4!}\delta\left(\varepsilon^{abcd}\varepsilon^{efgh}g_{ae}g_{bf}g_{cg}g_{dh}\right) = \frac{1}{6}\varepsilon^{abcd}\varepsilon^{efgh}g_{bf}g_{cg}g_{dh}\delta g_{ae}\\ &= \frac{1}{6}\varepsilon^{efgh}(g^{al}\bar{\varepsilon}_{lfgh}\det g)\delta g_{ae} = \det g g^{ae}\delta g_{ae}\\ &= -\det g\, g_{ae}\delta g^{ae} \end{aligned} \tag{3.40}$$

where we used $0 = \delta(g^{ae}g_{ae}) = g_{ae}\delta g^{ae} + g^{ae}\delta g_{ae}$ *in the last step. We conclude that*

$$\delta\sqrt{-\det g} = -\frac{\delta \det g}{2\sqrt{-\det g}} = -\frac{1}{2}\sqrt{-\det g}g_{ae}\delta g^{ae}\,. \tag{3.41}$$

With these variations, we finally have

$$\begin{aligned}\delta(\sqrt{-\det g}R_{ab}g^{ab}) &= -\frac{1}{2}\sqrt{-\det g}\, g_{cd}R_{ab}g^{ab}\delta g^{cd} + \sqrt{-\det g}R_{ab}\delta g^{ab}\\ &\quad + \sqrt{-\det g}\, g^{ab}\delta R_{ab}\\ &= \sqrt{-\det g}\left((R_{ab} - \tfrac{1}{2}Rg_{ab})\delta g^{ab} + g^{ab}\delta R_{ab}\right)\,. \end{aligned} \tag{3.42}$$

Here, $\sqrt{-\det g}g^{ab}\delta R_{ab} = \sqrt{-\det g}\nabla_a v^a = \partial_a(\sqrt{-\det g}v^a)$ *integrates to a boundary term by Stoke's theorem. Provided that there is no boundary or that boundary terms vanish by fall-off conditions,* $S_{\mathrm{EH}}[g]$ *is functionally differentiable, and its variation produces the Einstein tensor. We will discuss the issue of boundary terms in Chapter 3.3.2, once additional methods for dealing with hypersurfaces have been provided.*

3.2 Geometry of hypersurfaces

We now turn to the canonical formulation of gravity and apply the systematic procedures for an analysis of constrained systems. There will be several constraints with a non-trivial and very interesting algebra, and their expressions are rather involved, more so at least than the Gauss constraint of electromagnetism. It is thus helpful to perform the required calculations not by brute force, although a certain amount of force will be necessary, but by using geometrical techniques to split a space-time tensor into spatial parts.

3.2.1 Foliation

As S_{EH} in (3.36) is the correct action for Einstein's equation, it can be used as starting point for a canonical decomposition. A canonical formulation then requires a space-time splitting, since only time derivatives are transformed to momenta but not space derivatives. For field theories on Minkowski space, there are natural choices of time, and global Lorentzian symmetries exist to change one's choice of time. But we now face a situation in which no background symmetries or preferred (inertial) observers exist. The notion of time has become much more general. We thus only assume space-time M to be foliated by spatial surfaces $\Sigma \approx \Sigma_t$: $t =$ const of some time function t. Space-time is assumed to be globally hyperbolic: topologically, $M = \Sigma \times \mathbb{R}$.

3.2.1.1 Normal and spatial structures

A surface Σ_{t_0} in space-time can be considered as a constraint surface characterized by the constraint $C_{t_0} = t - t_0 = 0$ with a constant t_0. The difference from our previous constraints on phase space is, however, that we do not have a Poisson or symplectic geometry on space-time, but a Riemannian one given by the metric g_{ab} and its inverse g^{ab}. In lieu of a Hamiltonian vector field generated by the constraint, we have the vector $g^{ab}\partial_b C_{t_0} = g^{ab}\partial_b t$ obtained by raising the index of the 1-form $\mathrm{d}t$. (In analogy to the previous notation, we could write it as $g^\sharp \mathrm{d}t$.) In contrast to Poisson geometry, where the Hamiltonian vector field of a single, necessarily first-class constraint must be tangent to the constraint surface, Riemannian geometry ensures that $X^a = g^{ab}\partial_b t$ is normal to the surface: for any vector field s^a tangent to $t =$ const, we have $g_{ab}s^a X^b = s^a\partial_a t = 0$. (It is the antisymmetry of the Poisson tensor that makes the Hamiltonian vector field of a single constraint C tangent to the constraint surface: $X_C C = \mathcal{P}^{ij}\partial_j C \partial_i C = 0$.) The non-degeneracy of the metric also gives us a means to normalize the (timelike) normal vector to $n^a := X^a/\sqrt{-g_{bc}X^b X^c}$ such that $g_{ab}n^a n^b = -1$. (Obviously, Hamiltonian vector fields cannot be normalized in the context of Poisson geometry.) We finally identify a unique normal by requiring it to be future-pointing: $n^a\partial_a t > 0$. For globally hyperbolic space-times, this is always possible.

With this structure, the tangent space in TM at each point of Σ_t is naturally decomposed into a spatial tangent space spanned by vectors tangent to Σ_t and a normal space spanned by the unique unit future-pointing vector field n^a normal to Σ_t. Each spatial slice Σ_t is equipped with its own Riemannian structure: we have the *induced metric* $h_{ab} = g_{ab} + n_a n_b$ which is uniquely determined by the two conditions that $h_{ab}n^b = 0$ (it is insensitive to the normal direction) and $h_{ab}s^a = g_{ab}s^a$ for any vector s^a tangent to Σ_t (when applied to tangent vectors intrinsic to Σ_t, it gives the same geometry as g_{ab}). (Comparing again to Poisson geometry, the induced metric is analogous to the Dirac bracket, which subtracts from off the Poisson structure any contribution from the flow of constraints transversal to the constraint surface.) In the sense of its action on different vector fields, only spatial components are non-zero in h_{ab}, and they agree with those of g_{ab}. We can thus consider h_{ab} as the spatial part of g_{ab}, defining a positive definite metric on each spatial slice Σ_t. Similarly, we define $h^{ab} = g^{ab} + n^a n^b$. It reduces to the inverse of h_{ab} when applied to spatial vectors tangent to Σ_t, but is not the inverse of h_{ab} as a space-time tensor. In fact, neither h_{ab} nor h^{ab} is invertible on space-time.

3.2.1.2 Time derivatives

The field h_{ab} will play a crucial role as the configuration variable of canonical gravity. In order to define it, we had to make use of a time function, or of the foliation Σ_t it generates. Accordingly, we cannot view h_{ab} as a space-time tensor field, but rather, as a

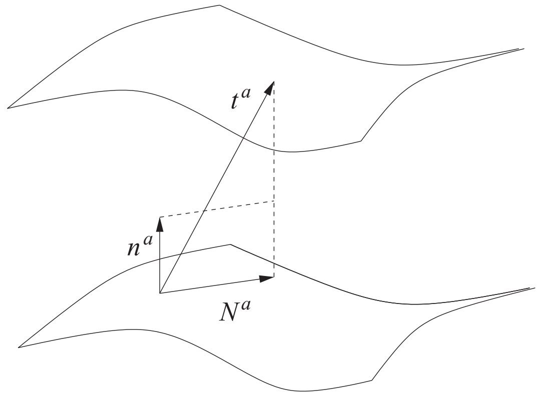

Fig. 3.3 Decomposition of a time-evolution vector field t^a along the normal n^a as $Nn^a + N^a$ with a spatial component N^a tangent to spatial slices.

4-dimensional tensor field on the foliated space-time M. Alternatively, we may interpret it as a time-dependent 3-dimensional tensor field on the family of manifolds Σ_t. This is the viewpoint we will take for the dynamics of the canonical formulation.

In this case, with a time-dependent interpretation of spatial fields, it makes sense to define time derivatives of the induced metric or other fields. A derivative requires a direction, which is not provided by the time function t solely made use of so far. In addition, we must introduce a *time-evolution vector field* t^a to define the direction of time derivatives, such that $t^a\nabla_a t = 1$ to ensure that it agrees with the sense of time provided by t. This condition ensures that $t^a\nabla_a$ can be interpreted as $\partial/\partial t$, i.e. as the partial derivative by time for some as yet unspecified spatial coordinates to be held fixed. (Notice that a partial time derivative $\partial/\partial t$ makes unique sense only if spatial coordinates x^b to be held fixed are known, such that $t^a\nabla_a x^b = 0$ in addition to $t^a\nabla_a t = 1$.)

If one does not intend to select spatial coordinates, the time-evolution vector field t^a is not uniquely defined even if the time function t is fixed. The freedom can be parameterized by decomposing t^a into a spatial part, the *shift vector* $N^a := h^{ab}t_b$, and the normal part $Nn^a = t^a - h^{ab}t_b$ with the *lapse function* $N = -n_a t^a$; see Fig. 3.3. In an isotropic space-time, for instance, we have that $N^a = 0$ as any non-vanishing spatial vector would break isotropy, and we can write $t^a\nabla_a = \partial/\partial t$ keeping co-moving spatial coordinates fixed. Thus, in FLRW models $n^a\nabla_a = N^{-1}\partial/\partial t$ is the proper-time derivative. Without isotropy, non-vanishing shift vectors are possible. One may interpret the unit-normal vector field n^a and the evolution vector field t^a as corresponding to the world-lines followed by two different families of observers, the so-called Eulerian observers following n^a, and ones in relative motion following t^a. Eulerian observers with 4-velocity field n^a are distinguished only by the geometry of the foliation, which may be chosen rather arbitrarily without considering a physical setting of interest; and so for sufficiently general physical situations to be described with any fixed foliation, the whole freedom in t^a must be allowed. For instance, to simplify calculations, the foliation may have been adapted to the flow of a matter fluid, but an observer should not be required to follow the same flow; t^a then describes the observer's motion with respect to the fluid.

As defined, t^a is not normalized and thus cannot directly be the 4-velocity of observers. If we normalize it, using the normalization of n^a and $n^a N_a = 0$, we obtain

$$\frac{t^a}{\sqrt{-||t||^2}} = \frac{n^a + N^a/N}{\sqrt{1 - (\vec{N}/N)^2}}$$

as the relativistic 4-velocity $(\gamma_{\vec{V}}, \gamma_{\vec{V}}\vec{V})$ expressed in the frame of Eulerian observers, with $\gamma_{\vec{V}} = (1 - \vec{V}^2)^{-1/2}$ corresponding to the 3-velocity $\vec{V} = \vec{N}/N$. The shift vector thus provides the velocity field relative to Eulerian observers.

Given a time-evolution vector field, we complete the interpretation of tensor fields on a foliated space-time as time-dependent tensor fields on space. To speak of the time-dependence of a field, we must be able to identify points at which we read off the time dependence. If we just have two different slices in the foliation, it is impossible to say how a field defined on them changes unless we can uniquely associate a point on one slice with a point on the other slice. Once this is available, evaluating the fields at the associated points shows their change when going from one slice to the next, that is their time dependence. A time-evolution vector field provides such an association: taking any one of its integral curves, we identify its intersections with all surfaces in Σ_t as corresponding to the same spatial point at different times. Thanks to the condition $t^a\nabla_a t = 1$, integral curves intersect each Σ_t exactly once, making the identification well defined. We can then meaningfully define a *time derivative* as the Lie derivative (as discussed in the Appendix) along the time-evolution vector field t^a:

$$\dot{T}^{a_1\cdots a_n}{}_{b_1\cdots b_m} := h^{a_1}_{c_1}\cdots h^{a_n}_{c_n} h^{d_1}_{b_1}\cdots h^{d_m}_{b_m} \mathcal{L}_t T^{c_1\cdots c_n}{}_{d_1\cdots d_m} \tag{3.43}$$

including spatial projections to ensure that $\dot{T}^{a_1\cdots a_n}{}_{b_1\cdots b_m}$ is spatial. (Without projections, $\mathcal{L}_t T^{a_1\cdots a_n}{}_{b_1\cdots b_m}$ is guaranteed to be spatial according to the foliation normal to n^a only if $[t,n]^a = 0$.)

3.2.1.3 Metric decomposition

Using the preceding definitions, we decompose the inverse space-time metric as

$$g^{ab} = h^{ab} - n^a n^b = h^{ab} - \frac{1}{N^2}(t^a - N^a)(t^b - N^b)\,. \tag{3.44}$$

Writing this as a matrix in a coordinate basis and inverting it, we obtain the line element

$$\mathrm{d}s^2 = -N^2\mathrm{d}t^2 + h_{ab}(\mathrm{d}x^a + N^a\mathrm{d}t)(\mathrm{d}x^b + N^b\mathrm{d}t) \tag{3.45}$$

in coordinates x^a such that $t^a\nabla_a = \partial/\partial t$. Space-time geometry is thus described not by a single metric but by the spatial geometry of slices, encoded in h_{ab}, together with deformations of neighboring slices with respect to each other as described by N and N^a.

For integration measures or other purposes, it is often important to know the determinant of the metric. For the space-time metric, its decomposition in spatial terms has a simple expression: from (3.44) we read off the inverse time-time component $g^{00} = -1/N^2$, while its purely spatial part is h^{ab}. The shift vector only contributes to the mixed time-space components. By a general formula for inverse matrices, the (i, j)-component of the inverse matrix of A is given by the co-factor C_{ji} (defined as $(-1)^{i+j}$ times the determinant of the submatrix obtained by striking out row j and column i) divided by the determinant of A: $(A^{-1})^{ij} = C_{ji}/\det A$. Here, for the $(0, 0)$-component of the inverse metric, we have the co-factor given by the determinant of h_{ab}, $C_{00} = \det h_{ab}$, and we already know that $g^{00} = -1/N^2 = C_{00}/\det(g_{cd})$. Thus,

$$\det(g_{cd}) = -N^2 \det(h_{ab})\,. \tag{3.46}$$

3.2.2 Intrinsic and extrinsic geometry

For a given foliation Σ_t of space-time, we define *spatial tensors* as those that vanish when contracted with the normal n^a on any index. Examples seen so far are the spatial metric h_{ab} and the shift vector N^a. Among spatial tensors, one can distinguish *intrinsic* and *extrinsic* geometrical notions, depending on whether only the spatial metric h_{ab} is used in the definition or also n^a separately, which refers to the slicing of space-time. Finally, there are tensors that depend even on the choice of time evolution vector field, such as N^a. Such tensors are not purely geometrical but require choices about the representation of dynamics.

3.2.2.1 Intrinsic geometry

The spatial metric itself is an intrinsic quantity, and as a metric it allows one to define a unique covariant derivative operator D_a on Σ such that $D_a h_{bc} = 0$. This covariant derivative can be written in terms of the space-time covariant derivative ∇_a as

$$D_c T^{a_1\ldots a_k}{}_{b_1\ldots b_l} = h^{a_1}{}_{d_1}\cdots h^{a_k}{}_{d_k} h_{b_1}{}^{e_1}\cdots h_{b_l}{}^{e_l} h_c{}^f \nabla_f T^{d_1\ldots d_k}{}_{e_1\ldots e_l}\,. \tag{3.47}$$

For a proof, it suffices to note that (3.47) satisfies all requirements of the unique covariant derivative: it immediately follows from the definition that it is linear, satisfies the Leibniz rule and preserves the spatial metric. Indeed,

$$h_a{}^d h_b{}^e h_c{}^f \nabla_f h_{de} = h_a{}^d h_b{}^e h_c{}^f \nabla_f (g_{de} + n_d n_e) = 0$$

since g_{de} is covariantly constant for ∇_f and in the second term at least one n_a will be contracted with h^{ab} after using the product rule. Even though the right-hand side of (3.47) does make use of space-time constructs, when applied to spatial tensors it is an intrinsic notion, since it is equivalent to the spatial covariant derivative defined solely by reference to h_{ab}.

Given the covariant derivative D_a, one defines *intrinsic-curvature tensors* as with any covariant derivative. For instance, we have the 3-dimensional Riemann tensor defined by

$$^{(3)}R_{abc}{}^d \omega_d = D_a D_b \omega_c - D_b D_a \omega_c \tag{3.48}$$

for all spatial ω_c, $\omega_a n^a = 0$. Since we make use only of the spatial derivative D_a, applied to spatial 1-forms, ${}^{(3)}R_{abc}{}^d$ is intrinsically defined. From the intrinsic Riemann curvature tensor, we obtain the Ricci tensor and scalar by the usual contractions.

3.2.2.2 Extrinsic curvature

The route to extrinsic geometrical notions is opened up by the observation that (3.47) is more general than the compatible covariant derivative of the spatial metric h_{ab}: it can be applied to any space-time tensor, not just spatial ones. After applying D_a, the result is always a spatial tensor according to (3.47), even if it is applied to a general space-time tensor. We thus obtain spatial tensors, but refer to space-time objects; the resulting notions of spatial geometry will be extrinsic.

Before introducing examples, we should be warned that (3.47) is not just an identity, but a definition of a more general derivative operator when $T^{a_1\ldots a_k}{}_{b_1\ldots b_l}$ is not spatial. The generalization sometimes provides pitfalls if one is not careful enough. In particular, when the covariant derivative is applied to a space-time tensor, one cannot use its compatibility with the spatial metric which, by construction, only applies when the derivative is taken of a purely spatial tensor. For instance, below in (3.49) we will consider the spatial tensor $D_a n_b$, with the unit normal n_a to spatial slices. Erroneously using compatibility, one would conclude that the trace $h^{ab} D_a n_b$ of this tensor must vanish because h^{ab} could be commuted with D_a, and $h^{ab} n_b = 0$. However, one cannot commute h^{ab} with D_a in this case because n_a is not a spatial tensor, and the general definition of D_a must be used; only the right-hand side of (3.47) can then be used as a definition of D_a, where h^{ab} and ∇_f do not commute.

In contrast to the intrinsic geometry, which applies to a single (Σ, h_{ab}) no matter how it is embedded in a space-time manifold, the *extrinsic geometry* of Σ in $\Sigma \times \mathbb{R}$ refers to the bending of Σ in its neighborhood, which in general implies a changing normal vector field n^a along Σ. This notion is captured in the definition of the *extrinsic-curvature tensor*

$$K_{ab} := D_a n_b = h_a^c h_b^d \nabla_c n_d \,. \tag{3.49}$$

It is a spatial tensor on Σ by definition of D_a, but is not intrinsically defined because we refer to n^a and thus to the embedding of Σ in space-time. Several properties of K_{ab} illustrate this tensor's meaning and will play a role later on:

1. $K_{ab} = h_a{}^c \nabla_c n_b = \nabla_a n_b + n_a n^c \nabla_c n_b$: we can drop one spatial metric from the general definition of D_a which would be contracted over the index b. This follows from

$$K_{ab} = h_b{}^d h_a{}^c \nabla_c n_d = (g_b{}^d + n_b n^d) h_a{}^c \nabla_c n_d = h_a{}^c \nabla_c n_b + n_b n^d h_a{}^c \nabla_c n_d$$

   noting that $n^d \nabla_c n_d = \frac{1}{2}(n^d \nabla_c n_d + n_d \nabla_c n^d) = \frac{1}{2} \nabla_c (n_d n^d) = 0$ (n^a is normalized). Here, we have several times commuted the space-time metric and its covariant derivative, which, in contrast to commuting h_{ab} with D_c, is always allowed.
2. $K_{ab} = K_{ba}$: the extrinsic curvature tensor is *symmetric*. To prove this, we take two arbitrary spatial tangent vector fields $Y^a \nabla_a = f^\alpha \partial/\partial x^\alpha$ and $Z^a \nabla_a = g^\alpha \partial/\partial x^\alpha$ and compute their commutator

$$[Y, Z]^a \nabla_a := \left(f^\alpha \frac{\partial g^\beta}{\partial x^\alpha} - g^\alpha \frac{\partial f^\beta}{\partial x^\alpha} \right) \frac{\partial}{\partial x^\beta} \,.$$

By definition, the commutator is also spatial. Thus, we have

$$0 = n_a[Y, Z]^a = n_a(Y^b\nabla_b Z^a - Z^b\nabla_b Y^a)$$
$$= -Z^a Y^b\nabla_b n_a + Y^a Z^b\nabla_b n_a = Y^a Z^b(\nabla_b n_a - \nabla_a n_b)$$

where we used $n_a Z^a = 0 = n_a Y^a$ before the last step. As a result, all spatial projections of $\nabla_a n_b$ are symmetric, which constitute K_{ab}. (This is a special case of the Frobenius theorem.)

3. Extrinsic curvature is (half) the Lie derivative of the intrinsic metric along the unit normal:

$$K_{ab} = \frac{1}{2}\mathcal{L}_n h_{ab} \tag{3.50}$$

To prove this, we compute

$$\mathcal{L}_n h_{ab} = n^c\nabla_c h_{ab} + h_{cb}\nabla_a n^c + h_{ac}\nabla_b n^c$$
$$= n^c\nabla_c(n_a n_b) + \nabla_a n_b + \nabla_b n_a$$
$$= (g_a{}^c + n_a n^c)\nabla_c n_b + (g_b{}^c + n_b n^c)\nabla_c n_a = 2K_{ab}$$

using the symmetry of K_{ab} in the last step.

4. Finally, we have

$$K_{ab} = \frac{1}{2N}(\dot{h}_{ab} - D_a N_b - D_b N_a) \tag{3.51}$$

where we use the time derivative $\dot{h}_{ab} = h_a{}^c h_b{}^d\mathcal{L}_t h_{cd}$ defined by evolution along t^a-trajectories as in (3.43). This follows from the last property:

$$K_{ab} = \frac{1}{2}\mathcal{L}_n h_{ab} = \frac{1}{2N}(Nn^c\nabla_c h_{ab} + h_{ac}\nabla_b(Nn^c) + h_{cb}\nabla_a(Nn^c))$$
$$= \frac{1}{2N}h_a{}^c h_b{}^d\mathcal{L}_{t-N}h_{cd} = \frac{1}{2N}h_a{}^c h_b{}^d(\mathcal{L}_t h_{cd} - \mathcal{L}_N h_{cd})$$

where we substituted $Nn^a = t^a - N^a$ and smuggled in projections $h_a^c h_b^d$, since we know that K_{ab} is spatial. The claimed identity then follows from the last step noting that $\mathcal{L}_N h_{cd}$, with the spatial shift vector N^a, can be computed purely spatially and equals $D_a N_b + D_b N_a$, since D_a is compatible with h_{ab}. (Notice that, according to our definitions, $\dot{h}_{ab}$ is not an object of either intrinsic or extrinsic geometry, since its definition requires not just knowledge of n^a or the foliation Σ_t, but also of the selection of a time-evolution vector field t^a. So do the lapse function N and the shift vector N^a. The dependence on properties of t^a not given solely by n^a cancels out in the combination with N and N^a on the right-hand side of (3.51).)

Example 3.8 (A 2-sphere in Euclidean 3-space)

Our definitions and calculations in this section were mostly insensitive to the signature of space-time, and so we can apply them in Euclidean space as well. A 2-sphere in Euclidean 3-space is most easily described in polar coordinates (r, ϑ, φ). *We have the line element* $\mathrm{d}s^2 = \mathrm{d}r^2 + r^2(\mathrm{d}\vartheta^2 + \sin^2\vartheta\,\mathrm{d}\varphi^2)$. *A surface at constant r has unit co-normal* $n_a = (\mathrm{d}r)_a$, *and carries the induced line element* $h_{ab}\mathrm{d}x^a\mathrm{d}x^b = (g_{ab} - n_a n_b)\mathrm{d}x^a\mathrm{d}x^b = r^2(\mathrm{d}\vartheta^2 + \sin^2\vartheta\,\mathrm{d}\varphi^2)$. *(Only one of the signs in this equation must be adapted to the signature, or more precisely the spacelike nature of the co-normal.) With constant components of the co-normal in the* *coordinates used,* $\nabla_a n_b = -\Gamma^r_{ab}$ *provides a diagonal matrix with components* $(0, r, r\sin^2\vartheta)$ *along the diagonal. Multiplying with* h_a^b*, which simply projects to the angular components,* K_{ab} *as a tensor on the sphere is represented by a diagonal matrix with components* $(r, r\sin^2\theta)$ *along the diagonal. This clearly agrees with the normal derivative of the metric according to (3.50), in this case* $K_{ab} = \frac{1}{2}\partial h_{ab}/\partial r$*. The trace of extrinsic curvature is* $K = 2/r$.

3.2.3 Curvature relations

Intrinsic and extrinsic curvature (3.48) and (3.49) together describe the space-time curvature, analogously to the splitting of the space-time metric g_{ab} into h_{ab} and n^a in (3.44). In the case of curvature, however, the correspondence is not complete as a simple counting of independent components shows: taking into account all its symmetries, the Riemann tensor in n dimensions has $n^2(n^2-1)/12$ independent components which results in 20 components of the space-time tensor and 6 for the spatial Riemann tensor. The extrinsic curvature tensor, which is symmetric, provides only 6 components more, which adds up to 12 rather than 20 independent quantities.

Nevertheless, what we have introduced so far captures all curvature components necessary for a canonical decomposition. The precise relation between curvature components is summarized in relations between the different tensors.

1. The *Gauss equation*

$$h_a{}^e h_b{}^f h_c{}^g h_h{}^d R_{efg}{}^h = {}^{(3)}R_{abc}{}^d + K_{ac}K_b{}^d - K_{bc}K_a{}^d \tag{3.52}$$

follows from

$$\begin{aligned} D_a D_b \omega_c &= D_a(h_b{}^d h_c{}^e \nabla_d \omega_e) = h_a{}^f h_b{}^g h_c{}^h \nabla_f (h_g{}^d h_h{}^e \nabla_d \omega_e) \\ &= h_a{}^f h_b{}^d h_c{}^e \nabla_f \nabla_d \omega_e + h_c{}^e (h_a{}^f h_b{}^g \nabla_f h_g{}^d) \nabla_d \omega_e + h_b{}^d (h_a{}^f h_c{}^h \nabla_f h_h{}^e) \nabla_d \omega_e \end{aligned}$$

where we have $h_a{}^f h_b{}^g \nabla_f h_g{}^d = h_a{}^f h_b{}^g \nabla_f (g_g{}^d + n_g n^d) = n^d h_b{}^g \nabla_a n_g = K_{ab} n^d$ in the second term and

$$\begin{aligned} h_b{}^d (h_a{}^f h_c{}^h \nabla_f h_h{}^e) \nabla_d \omega_e &= h_b{}^d K_{ac} n^e \nabla_d \omega_e = -K_{ac} h_b{}^d \omega_e \nabla_d n^e \\ &= -K_{ac} K_b^e \omega_e \end{aligned}$$

in the last term for spatial ω_e with $\omega_a n^a = 0$. Thus,

$$\begin{aligned} {}^{(3)}R_{abc}{}^e \omega_e &= D_a D_b \omega_c - D_b D_a \omega_c \\ &= h_a{}^f h_b{}^d h_c{}^e (\nabla_f \nabla_d \omega_e - \nabla_d \nabla_f \omega_e) - K_{ac} K_b^e \omega_e + K_{bc} K_a^e \omega_e \end{aligned}$$

which proves the identity.

2. *Codazzi equation:*

$$h^a{}_e h^b{}_f h^c{}_g R_{abcd} n^d = D_e K_{fg} - D_f K_{eg} \tag{3.53}$$

because the last term in

$$\begin{aligned}
h^a{}_e h^b{}_f h^c{}_g R_{abcd} n^d &= h^a{}_e h^b{}_f h^c{}_g (\nabla_a \nabla_b - \nabla_b \nabla_a) n_c \\
&= h^a{}_e h^b{}_f h^c{}_g (\nabla_a (g_b{}^d \nabla_d n_c) - \nabla_b (g_a{}^d \nabla_d n_c)) \\
&= D_e K_{fg} - h^a{}_e h^b{}_f h^c{}_g \nabla_a (n_b n^d \nabla_d n_c) - h^a{}_e h^b{}_f h^c{}_g \nabla_a (n_b n^d \nabla_d n_c) \\
&\quad - D_f K_{eg} + h^a{}_e h^b{}_f h^c{}_g \nabla_b (n_a n^d \nabla_d n_c) \\
&= D_e K_{fg} - D_f K_{eg} - h^a{}_e h^b{}_f h^c{}_g (n^d \nabla_d n_c)(\nabla_a n_b - \nabla_b n_a)
\end{aligned}$$

vanishes, using the symmetries of the spatial projection of $\nabla_a n_b$.

3. *Ricci equation:*

$$R_{acbd} n^c n^d = -\mathcal{L}_n K_{ab} + K_{ac} K^c_b + D_{(a} a_{b)} + a_a a_b \tag{3.54}$$

with the Lie derivative $\mathcal{L}_n$ along the unit normal n^a, and the normal acceleration $a_a := n^c \nabla_c n_a$ (satisfying $a_a n^a = 0$). This equation is slightly more tedious to derive than the others; we thus proceed in steps, sketching the main ingredients: first, we have

$$\mathcal{L}_n K_{ab} = n^c \nabla_c K_{ab} + K_{ac} \nabla_b n^c + K_{bc} \nabla_a n^c \,.$$

Using $K_{ab} = h^c_a \nabla_c n_b = \nabla_a n_b + n_a n^c \nabla_c n_b$, the first term becomes

$$\begin{aligned}
n^c \nabla_c K_{ab} &= n^c \nabla_c \nabla_a n_b + (n^c \nabla_c n_a)(n^d \nabla_d n_b) + n_a n^c \nabla_c (n^d \nabla_d n_b) \\
&= n^c \nabla_c \nabla_a n_b + a_a a_b - (\nabla_a n^d)(\nabla_d n_b) - n^d \nabla_a \nabla_d n_b + h^c_a \nabla_c (n^d \nabla_d n_b)\,.
\end{aligned}$$

In the last term, obtained after using $n_a n^c = -g^c_a + h^c_a$, we write

$$\begin{aligned}
h^c_a (h^e_b - n_b n^e) \nabla_c a_e &= D_a a_b - h^c_a n_b n^e \nabla_c a_e \\
&= D_a a_b + n_b a_e \nabla_a n^e + n_a n^c n_b a_e \nabla_c n^e \,.
\end{aligned}$$

Thus,

$$\begin{aligned}
n^c \nabla_c K_{ab} &= n^c (\nabla_c \nabla_a - \nabla_a \nabla_c) n_b + a_a a_b + D_a a_b - (\nabla_a n^d)(\nabla_d n_b) \\
&\quad + n_b (\nabla_a n^e)(n^c \nabla_c n_e) + n_a n_b (n^c \nabla_c n^e)(n^d \nabla_d n_e)\,.
\end{aligned}$$

The remaining terms in $\mathcal{L}_n K_{ab}$ can be written as

$$\begin{aligned}
K_{ac} \nabla_b n^c + K_{bc} \nabla_a n^c &= (\nabla_a n_c)(\nabla_b n^c) + n_b n^c (\nabla_c n^d)(\nabla_a n_d) + K_{ac}(K^c_b - n_b n^d \nabla_d n^c) \\
&= (\nabla_a n_c)(\nabla_b n^c) + K_{ac} K^c_b - n_a n^d (\nabla_d n_c) n_b n^e \nabla_e n^c \,.
\end{aligned}$$

Several of the terms contributing to $\mathcal{L}_n K_{ab}$ then cancel once we note that

$$\begin{aligned}
&(\nabla_a n_c)(\nabla_b n^c) - (\nabla_a n^d)(\nabla_d n_b) + n_b (\nabla_a n^e)(n^c \nabla_c n_e) \\
&= (\nabla_a n^c)\left(\nabla_b n_c - \nabla_c n_b + n_b n^d \nabla_d n_c\right) \\
&= (\nabla_a n^c)\left(h^d_b (\nabla_d n_c - \nabla_c n_d) + n_b n^d \nabla_c n_d\right) = 0
\end{aligned} \tag{3.55}$$

whose last term vanishes, since n_a is normalized, and the rest, since

$$h^d_b(\nabla_d n_c - \nabla_c n_d) = h^d_b h^e_c(\nabla_d n_e - \nabla_e n_d) - h^d_b n_c n^e(\nabla_d n_e - \nabla_e n_d)$$
$$= -n_c n^e \nabla_e n_b$$

is proportional to n_c, and vanishes when contracted with $\nabla_a n^c$ in (3.55). (The first terms in the last line vanish, since spatial projections of $\nabla_d n_e$ must form a symmetric tensor as seen in the context of extrinsic curvature.)

The Ricci equation then follows, since $R_{acbd}n^c n^d = n^c(\nabla_a\nabla_c - \nabla_c\nabla_a)n_b$.

4. It is useful to note a particular case of the Ricci equation that will directly be used below: there is a vector field v^a such that

$$R_{ab}n^a n^b = (K^a_a)^2 - K_a{}^b K_b{}^a + \nabla_a v^a\,. \tag{3.56}$$

This follows as a contraction of (3.54), or directly from

$$R_{ab}n^a n^b = R_{acb}{}^c n^a n^b = -n^a(\nabla_a\nabla_c - \nabla_c\nabla_a)n^c$$
$$= (\nabla_a n^a)(\nabla_c n^c) - (\nabla_c n^a)(\nabla_a n^c) - \nabla_a(n^a\nabla_c n^c) + \nabla_c(n^a\nabla_a n^c)$$

using

$$(\nabla_c n^a)(\nabla_a n^c) = g_a{}^d g^c{}_e(\nabla_c n^d)(\nabla_d n^c)$$
$$= (h_e{}^c\nabla_c n^a)(h_a{}^d\nabla_d n^e) - h_a{}^d n^c n_e(\nabla_c n^a)(\nabla_d n^c)$$
$$- n_a n^d h_e{}^c(\nabla_c n^a)(\nabla_d n^e) + n_a n^d n^c n_e(\nabla_c n^a)(\nabla_d n^c)$$
$$= K^a{}_e K^e{}_a$$

and an analogous calculation for $(\nabla_a n^a)(\nabla_c n^c)$. Thus, $v^a = -n^a\nabla_c n^c + n^c\nabla_c n^a$.

All other possible contractions of $R_{abc}{}^d$ with n^a or h_{ab} vanish due to symmetries of the Riemann tensor.

Example 3.9 (Foliated $1+1$-dimensional space-time)

In two space-time dimensions, the symmetries of the Riemann tensor require that it is fully determined by the Ricci scalar R such that $R_{abcd} = \frac{1}{2}R\epsilon_{ab}\epsilon_{cd}$ (with the totally antisymmetric $\epsilon_{ab} = \sqrt{|\det g|}\varepsilon_{ab}$ without density weight). A 1-dimensional spatial slice Σ has, up to orientation, a unique unit tangent vector s^a in addition to the usual future-pointing unit normal. We fix the orientation by choosing $s_a = \epsilon_{ab}n^b$ (as opposed to $-\epsilon_{ab}n^b$) which is guaranteed to be normalized and orthogonal to n^a. Only symmetric spatial tensors exist, and for degree n are proportional to $s_{a_1}\cdots s_{a_n}$. The spatial metric, for instance, is $h_{ab} = s_a s_b$ with a unit prefactor following from the normalization; $h^{ab}h_{ab} = (s^a s_a)^2 = 1$ as required for a projector to a 1-dimensional space. Similarly, there is a scalar K such that $K_{ab} = Kh_{ab}$.

The Gauss and Codazzi equations are trivially satisfied in this situation, since they would require non-zero spatial antisymmetric 2-tensors which do not exist here. (For instance, $\epsilon_{cd}h^c_a h^d_b = 0$.) The Ricci equation is more interesting: first, we have $R_{acbd}n^c n^d = \frac{1}{2}R\epsilon_{ac}n^c\epsilon_{bd}n^d = \frac{1}{2}Rh_{ab}$ using $\epsilon_{ab}n^b = s_a$. On the right-hand side of the Ricci equation,

$\mathcal{L}_n K_{ab} = \mathcal{L}_n(K h_{ab}) = h_{ab}\mathcal{L}_n K + 2K^2 h_{ab}$. The acceleration vector satisfies $a_a = a s_a$ with a scalar a. Here, $D_a a_b = s_a s^c s_b s^d \nabla_c(a s_d) = s_a s_b \left(a s^c s^d \nabla_c s_d + s^c \nabla_c a\right) = h_{ab}\mathcal{L}_s a$. For the scalars multiplying h_{ab}, the Ricci equation then becomes

$$\frac{1}{2}R = -K^2 - \mathcal{L}_n K + a^2 + \mathcal{L}_s a\,. \tag{3.57}$$

3.3 ADM formulation of general relativity

Extrinsic curvature, as shown by (3.51), plays the role of a "velocity" of the spatial metric and is thus a candidate for its momentum. It indeed appears in the Einstein–Hilbert action due to the Gauss–Codazzi relations: using symmetries of the Riemann tensor, we have

$$\begin{aligned} R = (h^{ab} - n^a n^b)(h^{cd} - n^c n^d)R_{abcd} &= h^{ab}h^{cd}R_{abcd} - 2R_{ab}n^a n^b \\ &= h^{ef}h^{gh}h_e{}^a h_f{}^b h_g{}^c h_h{}^d R_{abcd} - 2R_{ab}n^a n^b \\ &= {}^{(3)}R + K_{ab}K^{ab} - (K^a{}_a)^2 - 2\nabla_a v^a\,. \end{aligned}$$

Up to the divergence $\nabla_a v^a$, we can thus decompose the Ricci scalar into a "kinetic" term quadratic in extrinsic curvature and a "potential" term ${}^{(3)}R$ which only depends on the spatial metric and its spatial derivatives.

The Lagrangian, with $\det g = -N^2 \det h$ from (3.46), becomes

$$\begin{aligned} L_{\text{grav}} &= \frac{1}{16\pi G}\int \mathrm{d}^3x\sqrt{-\det g}\; R \\ &= \frac{1}{16\pi G}\int \mathrm{d}^3x N\sqrt{\det h}({}^{(3)}R + K_{ab}K^{ab} - (K_a{}^a)^2) \end{aligned} \tag{3.58}$$

up to boundary terms which do not affect local field equations. (Boundary terms will be discussed shortly.)

3.3.1 Constraints

From (3.58) we immediately see that the action depends on $\dot{h}_{ab}$ via K_{ab} but, as expected from the considerations at the beginning of this chapter, is independent of time derivatives of the remaining space-time metric components N and N^a. We thus have primary constraints,

$$p_N(x) = \frac{\delta L_{\text{grav}}}{\delta \dot{N}(x)} = 0\,, \quad p_a(x) = \frac{\delta L_{\text{grav}}}{\delta \dot{N}^a(x)} = 0\,. \tag{3.59}$$

These are all the primary constraints because the relation

$$p^{ab}(x) = \frac{\delta L_{\text{grav}}}{\delta \dot{h}_{ab}(x)} = \frac{1}{2N}\frac{\delta L_{\text{grav}}}{\delta K_{ab}} = \frac{\sqrt{\det h}}{16\pi G}(K^{ab} - K^c_c h^{ab}) \tag{3.60}$$

can be inverted for

$$\dot{h}_{ab} = \frac{16\pi G N}{\sqrt{\det h}}(2p_{ab} - p_c^c h_{ab}) + 2D_{(a}N_{b)}\,. \tag{3.61}$$

We have the total gravitational Hamiltonian

$$\begin{aligned} H_{\text{grav}} &= \int \mathrm{d}^3x \left(\dot{h}_{ab}p^{ab} - L_{\text{grav}} + \lambda p_N + \mu^a p_a\right) \\ &= \int \mathrm{d}^3x \left(\frac{16\pi G N}{\sqrt{\det h}}\left(p_{ab}p^{ab} - \frac{1}{2}(p_c^c)^2\right) + 2p^{ab}D_aN_b - \frac{N\sqrt{\det h}}{16\pi G}{}^{(3)}R\right) \\ &\quad + \int \mathrm{d}^3x(\lambda p_N + \mu^a p_a)\,. \end{aligned} \tag{3.62}$$

The primary constraints imply secondary constraints

$$0 = \dot{p}_N = \{p_N, H_{\text{grav}}\} = -\frac{16\pi G}{\sqrt{\det h}}\left(p_{ab}p^{ab} - \frac{1}{2}(p_a^a)^2\right) + \frac{\sqrt{\det h}}{16\pi G}{}^{(3)}R =: -C_{\text{grav}} \tag{3.63}$$

called the *Hamiltonian constraint*, and

$$0 = \dot{p}_{N^a} = \{p_{N^a}, H_{\text{grav}}\} = 2\sqrt{\det h}D^b\left((\det h)^{-1/2}p_{ab}\right) = 2D_b p_a^b =: -C_a^{\text{grav}} \tag{3.64}$$

called the *diffeomorphism constraint*. To derive C_a^{grav}, we have integrated by parts, using $2\int \mathrm{d}^3x\sqrt{\det h}D_a(p^{ab}N_b/\sqrt{\det h})$ as a boundary term for any vector field N^a. Also, this boundary term is ignored for now, but will play an important role later.

With these definitions, we see that the total Hamiltonian (3.62) is a linear combination of constraints (times multipliers):

$$H_{\text{grav}} = \int \mathrm{d}^3x(NC_{\text{grav}} + N^aC_a^{\text{grav}} + \lambda p_N + \mu^a p_{N^a}) + H_{\partial\Sigma} \tag{3.65}$$

up to boundary terms $H_{\partial\Sigma}$ whose form we will determine in the next section. The lapse function N and shift vector N^a now play the role of Lagrange multipliers of secondary constraints. There is no proper Hamiltonian which would be non-trivial on the constraint surface. This is in agreement with the fact that there is no absolute time in general relativity, since a non-vanishing Hamiltonian would generate time evolution in an external time parameter. (Only non-vanishing boundary terms based on fixed boundary values for the metric could lead to unconstrained evolution with respect to a preferred boundary time.) Instead, dynamics is determined by the constraints, such that evolution as a gauge flow can be parameterized arbitrarily. In this way, we see the reparameterization invariance of coordinates in a generally covariant theory. In fact, explicit calculations (to be discussed in more detail in Chapter 3.3.4) confirm that the constraints are first class and thus generate gauge transformations which do not change the physical information in solutions. The Hamiltonian constraint does this for time, and the diffeomorphism constraint for spatial coordinates. Once these constraints are satisfied, we make sure that the formulation is space-time covariant even though we started the canonical formulation with the slicing of space-time determined by a time function t.

3.3.2 Boundary terms

We have so far ignored boundary terms, in the variation of the Einstein–Hilbert action in Example 3.7 as well as in variations done so far for the canonical formulation. If space is without boundaries, boundary terms clearly play no role, and in other cases they can sometimes vanish by boundary conditions. However, not all situations in general relativity allow sufficiently strong fall-off conditions to make all boundary terms vanish, since the metric must remain non-degenerate; not all its components can drop off to zero.

A discussion of boundary terms is required to ensure that the Lagrangian varied is indeed functionally differentiable by the fields. Functional differentiability means that we can write the variation of the Lagrangian $L_{\rm grav}$ as

$$\delta L_{\rm grav} = \int \mathrm{d}^3x \left(A^{ab}\delta h_{ab} + B^{ab}\delta K_{ab}\right) \tag{3.66}$$

with two functionals A^{ab} and B^{ab} which we then identify with $\delta L_{\rm grav}/\delta h_{ab}$ and $\delta L_{\rm grav}/\delta K_{ab}$, respectively. If there is a boundary and boundary terms do not vanish by virtue of fall-off conditions for the fields or restrictions on the variations, the varied Lagrangian is not of the form (3.66) with only a bulk contribution, and functional derivatives are undefined. To ensure functional differentiability, boundary terms must be added to the original action, in order to cancel those terms arising from integrating the variation by parts. The values of boundary terms will also play an important physical role in Chapters 4.2.2 and 5.3.6.

3.3.2.1 Gibbons–Hawking term

In Example 3.7 we ignored the term $g^{ab}\delta R_{ab}$ when verifying that the Einstein–Hilbert action has extrema given by solutions to Einstein's equation. Thanks to Eq. (3.38), this term is the total divergence

$$g^{ab}\delta R_{ab} = -\nabla_a(g^{ab}\delta C^c_{cb} - g^{bc}\delta C^a_{bc})$$

which does not contribute to the local field equations if the action is functionally differentiable. To ensure that all boundary terms cancel, we now compute the value of the resulting contribution.

As always, in order to fully specify the variational principle used, we must determine how the fields are allowed to be varied at the boundary. In classical mechanics, for instance, Euler–Lagrange equations for a particle's motion are found by varying the trajectory $q(t)$, keeping its endpoints fixed. While δq is unrestricted at all times between the initial time $t_{\rm initial}$ and the final one, $t_{\rm final}$, $\delta q(t_{\rm initial})$ and $\delta q(t_{\rm final})$ are held constant. (After having determined the equations of motion, it is sometimes useful to allow more general variations, for instance to find conserved quantities.) Similarly, variations of fields on manifolds with boundaries of space, amounting to timelike boundaries in space-time, cannot be completely free at the boundaries. In what follows, we will require the induced metric to be fixed at the boundary, while the space-time metric in a neighborhood, and thus extrinsic curvature of the boundary, may vary. To be specific, we will assume the boundary to be timelike; a fixed boundary metric can then be interpreted as an asymptotic reference frame carried by a family of observers moving along the boundary. To avoid convergence issues, we assume the boundary to be at a finite location in space; truly asymptotic properties will be introduced and used for boundary observables in Chapter 5.3.6.

Using the formula (3.37) for the varied connection, we obtain $g^{ab}\delta C^c_{cb} = \frac{1}{2}g^{cd}\nabla^a\delta g_{cd}$ and $g^{bc}\delta C^a_{bc} = g^{bc}g^{ad}\nabla_b\delta g_{cd} - \frac{1}{2}g^{bc}\nabla^a\delta g_{bc}$, easily combined to provide the total-divergence term $g^{ab}\delta R_{ab} = \nabla^a(\nabla_b\delta g_{ac} - \nabla_a\delta g_{bc})g^{bc}$. Thus, if space-time M has a boundary ∂M, a boundary term

$$\begin{aligned}\int_M \mathrm{d}^4x\sqrt{-\det g}\; g^{ab}\delta R_{ab} &= \int_{\partial M}\mathrm{d}^3y\sqrt{-\det q}\; r^a g^{bc}(\nabla_b\delta g_{ac} - \nabla_a\delta g_{bc})\\ &= \int_{\partial M}\mathrm{d}^3y\sqrt{-\det q}\; r^a q^{bc}(\nabla_b\delta g_{ac} - \nabla_a\delta g_{bc})\end{aligned}$$

arises. Here, we denote coordinates on the boundary by y. For a timelike boundary, as assumed, the induced metric q_{ab} has negative determinant. The spacelike unit normal to the boundary is the vector field r^a. Thus, $q^{ab} = g^{ab} - r^a r^b$ by a change of signature compared to our induced metrics for spacelike slices. We have made use of this relationship as well as of antisymmetry properties in the last step.

Let us now restrict the metric variations to those leaving the boundary metric and the normal vector field unchanged: $\delta g_{ab}|_{\partial M} = 0$. This eliminates the first term of the previous result, since $q^{bc}\nabla_b$ is a derivative tangent to the boundary, along which δg_{ac} is constant. We are left with

$$\int_M \mathrm{d}^4x\sqrt{-\det g}\; g^{ab}\delta R_{ab} = -\int_{\partial M}\mathrm{d}^3y\sqrt{-\det q}\; r^a q^{bc}\nabla_a\delta g_{bc}\,. \tag{3.67}$$

In this expression, a normal derivative $r^a\nabla_a$ of the metric features, which looks similar to Eq. (3.50) obtained for spacelike surfaces. Indeed, it is easy to relate the remaining variation to one of extrinsic curvature of the boundary, more precisely the trace of extrinsic curvature $K = q^{ab}\nabla_b r_a$. Since the induced metric and the normal are held fixed, the only term contributing to δK is the connection Γ^c_{ab}, from which we have

$$\delta K = -q^{ab}\delta C^c_{ab} r_c = \frac{1}{2}q^{ab}r^c\nabla_c\delta g_{ab}\,.$$

(The other terms in δC^c_{ab} have only derivatives of δg_{ab} tangent to the boundary.) The result is exactly what we have in (3.67), and we write

$$\int_M \mathrm{d}^4x\sqrt{-\det g}\; g^{ab}\delta R_{ab} = -2\int_{\partial M}\mathrm{d}^3y\sqrt{-\det q}\;\delta K\,. \tag{3.68}$$

The boundary term vanishes only if δK vanishes. However, we have already decided that the induced boundary metric remains unchanged under variations; also to fix the extrinsic curvature would usually be too restrictive. There may thus be a non-vanishing boundary term — unless we amend the original action, S_{EH}, by a boundary term to cancel the one produced

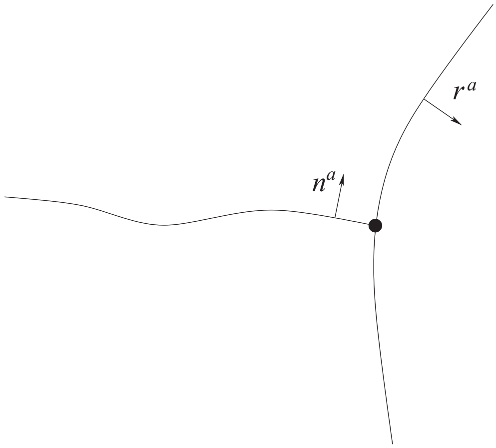

Fig. 3.4 A spatial slice with normal n^a, intersecting the timelike space-time boundary with outward-pointing normal r^a.

here. This is indeed possible, since (3.68) is the total variation $-2\delta\left(\int_{\partial M} \mathrm{d}^3y\sqrt{-\det q}\,K\right)$. If we use the action

$$S = S_{\mathrm{EH}} + \frac{1}{8\pi G}\int_{\partial M} \mathrm{d}^3y\sqrt{-\det q}\,K \tag{3.69}$$

amended by the Gibbons–Hawking boundary term (as proposed by Gibbons and Hawking (1977)), all boundary terms in variations cancel and the action is functionally differentiable in the presence of boundaries with fixed induced metric. A detailed account of boundary terms in variations of the Einstein–Hilbert action has been given by York (1986).

3.3.2.2 Boundary contribution to constraints

With the Gibbons–Hawking term, the action is functionally differentiable by metric variations fixed at the boundary. The boundary term must then also be included in the canonical decomposition, where it will provide extra terms. Returning now to the canonical analysis, we notice that another boundary term was ignored when writing the Einstein–Hilbert action in the form (3.58), producing a combination of different boundary terms. We already know the form of the Gibbons–Hawking term; now we can compute the remaining one to see how it all combines.

In the situation of constraints in the presence of boundaries, we are led to consider two different 3-dimensional submanifolds of space-time: the boundary, which we still assume to be timelike, and a spatial slice; see Fig. 3.4. The boundary has unit normal r^a, and the induced metric $q^{ab} = g^{ab} - r^a r^b$ as before. The spatial slice Σ with its unit normal n^a carries the canonical induced metric $h^{ab} = g^{ab} + n^a n^b$. At the boundary $\partial\Sigma$ of the spatial slice these two submanifolds intersect in a 2-dimensional surface. On that surface we can induce a metric σ_{ab} in two ways: from the spatial slice or from the boundary. We obtain $\sigma^{ab} = h^{ab} - r^a r^b = q^{ab} + n^a n^b$. Here and in what follows we make the simplifying assumption that $n^a r_a = 0$. Thus, r^a is tangent to Σ and provides the normal to $\partial\Sigma$ within Σ; similarly, n^a is tangent to ∂M and provides the normal to $\partial\Sigma$ within ∂M.

When expressing the 4-dimensional Ricci scalar in terms of the 3-dimensional one and extrinsic curvature by

$$R = {}^{(3)}R + K^{ab}K_{ab} - K^2 - 2\nabla_a(n^b\nabla_b n^a - n^a\nabla_b n^b)$$

we had so far ignored the total divergence. If the spatial slice Σ has a boundary $\partial\Sigma$, the Lagrangian (3.58) will receive a boundary term from this divergence, in addition to the Gibbons–Hawking term restricted to $\partial\Sigma$. Since the total divergence is obtained for a space-time vector field, we have to start from the space-time action and work our way down to the 2-surface $\partial\Sigma$. Using the normal r^a to the boundary ∂M and the induced metric q_{ab}, we have

$$\begin{aligned}-2\int_M \mathrm{d}^4x\sqrt{-\det g}\,\nabla_a v^a &= -2\int_{\partial M}\mathrm{d}^3y\sqrt{-\det q}\,r_a(n^b\nabla_b n^a - n^a\nabla_b n^b)\\ &= -2\int_{\partial M}\mathrm{d}^3y\sqrt{-\det q}\,r_a n^b\nabla_b n^a\end{aligned}$$

thanks to the assumption $r_a n^a = 0$.

On the boundary, the cross-section $\partial\Sigma$ plays the role of a spatial slice in the 3-dimensional boundary space-time ∂M; its metric determinant satisfies $N\sqrt{\det\sigma} = \sqrt{-\det q}$ with the lapse function N at the boundary. Moreover, we denote coordinates on $\partial\Sigma$ by z, which combines with the time coordinate t to the coordinates y on ∂M; in particular, $\mathrm{d}^3y = \mathrm{d}^2z\mathrm{d}t$. Then, the purely spatial boundary contribution to the constraints, splitting off the t-integration, is

$$-2\int_{\partial\Sigma}\mathrm{d}^2zN\sqrt{\det\sigma}\,r_a n^b\nabla_b n^a = 2\int_{\partial\Sigma}\mathrm{d}^2zN\sqrt{\det\sigma}\,n^a n^b\nabla_b r_a$$

using again our simplifying assumption $n^a r_a = 0$.

This term is to be combined with the restriction of the Gibbons–Hawking term to $\partial\Sigma$, resulting in

$$\begin{aligned}2\int_{\partial\Sigma}\mathrm{d}^2zN\sqrt{\det\sigma}\,(K + n^a n^b\nabla_b r_a) &= 2\int_{\partial\Sigma}\mathrm{d}^2zN\sqrt{\det\sigma}\,(q^{ab}+n^a n^b)\nabla_b r_a\\ &= 2\int_{\partial\Sigma}\mathrm{d}^2zN\sqrt{\det\sigma}\,\sigma^{ab}\nabla_b r_a\\ &= 2\int_{\partial\Sigma}\mathrm{d}^2zN\sqrt{\det\sigma}\,k\end{aligned}$$

as the boundary term of the canonical Lagrangian (3.58). All terms combine to the trace of extrinsic curvature k, now corresponding to the spatial boundary $\partial\Sigma$ as a surface in Σ with unit normal r^a.

Finally, there is a boundary term $2\int \mathrm{d}^2z\sqrt{\det\sigma}\,r_a(p^{ab}N_b/\sqrt{\det h})$ that arose when we computed the diffeomorphism constraint by varying the shift vector. The total Hamiltonian thus receives a boundary contribution

$$H_{\partial\Sigma}[N, N^a] = \int_{\partial\Sigma} \mathrm{d}^2 z \left(-(8\pi G)^{-1} \sqrt{\det \sigma}\, Nk + 2 r_a p^{ab} N_b / N \right) \tag{3.70}$$

as derived by Brown and York (1993). See also the reformulation by Booth and Fairhurst (2003) using phase-space techniques on the solution space of general relativity. (Although the timelike boundary of space-time no longer explicitly enters in this formula, which, rather, refers only to a spatial slice and its own boundary, we have used the orthogonality of the normals n^a and r^a in the derivation. The equation, thus, strictly applies only in the case of a foliation of space-time for which $\partial\Sigma_t$ forms a timelike boundary orthogonal to the spatial slices Σ_t. The general case has been discussed from different perspectives by Hayward (1993), Hawking and Hunter (1996) and Booth and Mann (1999).)

This expression plays an important role in the definition of *quasilocal quantities* of energy and momentum ("quasilocal" in the sense that it refers to submanifolds in space-time — the boundary — rather than single points). A direct application, however, would give unacceptable results. For instance, if we take part of a spatial slice $t = \text{const}$ in Minkowski space, bounded by a 2-sphere of radius R, the boundary Hamiltonian is non-vanishing: According to Example 3.8, the sphere has a trace of extrinsic curvature $k = 2/R$, such that a slicing with $N = 1$, $N^a = 0$, corresponding to a static observer, provides a non-zero Hamiltonian $H_{\partial\Sigma}[1, \vec{0}] = -R/G$. The conditions of orthogonality of spatial slices and the timelike boundary in space-time are met, and there are no additional terms in the boundary Hamiltonian. On the other hand, Minkowski space is empty and should not provide a non-zero result for energies. To avoid the conclusion that we have to assign non-zero energies to regions of Minkowski space even as measured by static observers, we "normalize" the boundary Hamiltonian (3.70) by subtracting from it its value obtained for a reference space-time whose energy behavior we think we understand.[6] In particular, we expect energy and momentum to vanish for Minkowski space-time, justifying the normalization

$$H^{\mathrm{norm}}_{\partial\Sigma}[N, N^a] = \int_{\partial\Sigma} \mathrm{d}^2 z \Big(-N(8\pi G)^{-1} (\sqrt{\det\sigma}\, k - \sqrt{\det\bar{\sigma}}\, \bar{k}) + 2N_b (r_a p^{ab} - \bar{r}_a \bar{p}^{ab})/N \Big) \tag{3.71}$$

in which all barred quantities refer the tensors computed for our reference space-time. The Minkowski quantities then obviously vanish.

If we consider a space-time other than the Minkowski one, non-zero values of $H^{\mathrm{norm}}_{\partial\Sigma}$ will signal deviations from the Minkowski behavior related to the Hamiltonian and thus to energies. In particular, if that space-time allows a timelike Killing vector field ξ^a, which also leaves the reference space-time invariant, and if we use ξ^a as our time-evolution vector field $t^a = Nn^a + N^a$, evolution of the canonical boundary data is generated by the Hamiltonian

[6] Such subtractions are required especially if asymptotic boundary quantities are to be computed by taking a limit in which the boundary approaches infinity. Without subtractions, the boundary integral would diverge as seen by the Minkowski example. We will discuss asymptotic properties in a later chapter.

$H_{\rm grav}[N, N^a] = H^{\rm norm}_{\partial\Sigma}[N, N^a]$ since all the constraints in the bulk term must be satisfied. For a Killing vector field ξ^a, this is a conserved quantity whose components,

$$E = -(8\pi G)^{-1} \int_{\partial\Sigma} {\rm d}^2 z \left(N(\sqrt{\det\sigma}\, k - \sqrt{\det\bar\sigma}\, \bar k)\right) \tag{3.72}$$

and

$$J = 2 \int_{\partial\Sigma} {\rm d}^2 z \left(N_b(r_a p^{ab} - \bar r_a \bar p^{ab})/N\right) \tag{3.73}$$

including the subtraction for normalization, are the Brown–York quasilocal energy E and (angular) momentum J (depending on whether N^a is a rotational or translational vector field).

Example 3.10 (Quasilocal energy in the Schwarzschild space-time)
In the Schwarzschild space-time described by the original Schwarzschild coordinates, families of spheres at constant $r = R$ provide timelike boundaries orthogonal to spatial slices $t = \text{const}$. Their unit normal vector field is obtained from $r_a = ({\rm d}r)_a/\sqrt{1-2GM/R}$. Extrinsic curvature, following Example 3.8 but using the Schwarzschild connection, leads to a diagonal extrinsic-curvature tensor with components $(R\sqrt{1-2GM/R}, R\sqrt{1-2GM/R}\sin^2\vartheta)$ along the diagonal. Its trace is $k = 2\sqrt{1-2GM/R}/R$. The quasilocal energy for static observers along these spheres is $H^{\rm norm}_{\partial\Sigma}[1,\vec 0] = -(\sqrt{1-2GM/R} - 1)R/G$ which is finite and, as expected, provides a formula for the black-hole mass. If $R \gg 2GM$, $H^{\rm norm}_{\partial\Sigma}[1,\vec 0] = M(1 + O(GM/R))$.

3.3.3 Equations of motion

The variational problem with the correct boundary term is well defined, and we can proceed to compute local equations of motion generated by the Hamiltonian. Hamiltonian equations of motion of the general type (3.15), (3.16) then give $\dot N(x) = \lambda(x)$ and $\dot N^a(x) = \mu^a(x)$, which tells us that these functions can change arbitrarily due to reparameterizations. Moreover, $\dot h_{ab} = \{h_{ab}, H_{\rm grav}\}$ just reproduces the equation (3.61) in terms of the momentum. Finally, we obtain a non-trivial evolution equation

$$\dot p^{ab} = \{p^{ab}, H_{\rm grav}\} = -\frac{\delta H_{\rm grav}}{\delta h_{ab}}$$

which we compute in several steps, using the basic expression

$$\frac{\delta h_{cd}(x)}{\delta h_{ab}(y)} = \delta^a{}_{(c}\delta^b{}_{d)}\delta(x,y) \tag{3.74}$$

of the functional derivative as a symmetrized product of delta tensors and functions.

We denote the different h_{ab}-dependent terms in $H_{\rm grav}$ as

$$H_1 := \int \mathrm{d}^3x \frac{16\pi GN}{\sqrt{\det h}} \left(p^{ab} p^{cd} h_{ac} h_{bd} - \frac{1}{2} p^{ab} p^{cd} h_{ab} h_{cd} \right) \tag{3.75}$$

$$H_2 := 2 \int \mathrm{d}^3x p^{ab} D_a N_b = 2 \int \mathrm{d}^3x p^{ab} h_{bc} (\partial_a N^c + G^c_{ab} N^d) \tag{3.76}$$

$$H_3 := - \int \mathrm{d}^3x \frac{N\sqrt{\det h}}{16\pi G} h^{ab}\, {}^{(3)}R_{ab} \tag{3.77}$$

where G^c_{ab} are connection coefficients for the spatial covariant derivative D_a. In these expressions, it is important to use indices in positions as they appear in the basic definitions, for instance p^{ab} with upper indices. (This tensor density is canonically conjugate to h_{ab} and thus to be held fixed in functional derivatives by h_{ab}.) Otherwise, one could overlook factors of the spatial metric which contribute to the functional derivatives. Similarly, covariant derivatives are metric-dependent via the connection coefficients and provide non-vanishing functional derivatives.

In the variations of H_1, H_2 and H_3 we will make use of of the formulas (3.37) and (3.41) already computed in Example 3.7:

$$\delta \det h = \det h \; h^{ab} \delta h_{ab} \tag{3.78}$$

$$\delta G^a_{bc} = \frac{1}{2} h^{ad} (D_b \delta h_{cd} + D_c \delta h_{bd} - D_d \delta h_{bc}) . \tag{3.79}$$

These relations also hold in the three spatial dimensions used here, not just in four space-time dimensions for which they were derived explicitly. We then have

$$\frac{\delta H_1}{\delta h_{ab}} = \frac{32\pi GN}{\sqrt{\det h}} \left(p^{ac} p^{bd} h_{cd} - \frac{1}{2} p^{ab} p^c{}_c \right) - \frac{8\pi GN}{\sqrt{\det h}} h^{ab} \left(p^{cd} p_{cd} - \frac{1}{2} (p^c{}_c)^2 \right)$$

from the explicit metric factors as well as the determinant,

$$\begin{aligned}
\frac{\delta H_2}{\delta h_{ab}} &= 2p^{(a|c|} D_c N^{b)} + 2 \int \mathrm{d}^3x p^{cd} h_{de} \frac{\delta G^e_{cf}}{\delta h_{ab}} N^f \\
&= 2p^{c(a} D_c N^{b)} + \int \mathrm{d}^3x \left(p^{cd} N^f \left(D_c \left(\delta^a{}_{(f} \delta^b{}_{d)} \delta(x,y) \right) \right.\right. \\
&\quad \left.\left. + D_f \left(\delta^a{}_{(c} \delta^b{}_{d)} \delta(x,y) \right) - D_d \left(\delta^a{}_{(c} \delta^b{}_{f)} \delta(x,y) \right) \right) \right) \\
&= 2p^{c(a} D_c N^{b)} - \sqrt{\det h} \left(D_c \left(p^{c(b} N^{a)} / \sqrt{\det h} \right) \right. \\
&\quad \left. + D_f \left(p^{ab} N^f / \sqrt{\det h} \right) - D_d \left(p^{(a|d|} N^{b)} / \sqrt{\det h} \right) \right) \\
&= 2p^{c(a} D_c N^{b)} - \sqrt{\det h} D_c \left(p^{ab} N^c / \sqrt{\det h} \right)
\end{aligned}$$

with several terms from the derivative of connection coefficients, and finally

$$\begin{aligned}\frac{\delta H_3}{\delta h_{ab}} &= \frac{N\sqrt{\det h}}{16\pi G}{}^{(3)}R^{ab} - \frac{N\sqrt{\det h}}{32\pi G}h^{ab(3)}R + \int \mathrm{d}^3x \frac{N\sqrt{\det h}}{16\pi G}\left(D^c\frac{\delta G^d_{dc}}{\delta h_{ab}} - D_d\frac{\delta G^{d}{}_{c}{}^{c}}{\delta h_{ab}}\right)\\ &= \frac{N\sqrt{\det h}}{16\pi G}{}^{(3)}R^{ab} - \frac{N\sqrt{\det h}}{32\pi G}h^{ab(3)}R + \sqrt{\det h}\left(h^{ab}D_cD^cN - D^aD^bN\right) .\end{aligned}$$

In the last calculation, we have used

$$\begin{aligned}&\int \mathrm{d}^3x \frac{N\sqrt{\det h}}{16\pi G}\left(D^c\frac{\delta G^d_{dc}}{\delta h_{ab}} - D_d\frac{\delta G^{d}{}_{c}{}^{c}}{\delta h_{ab}}\right)\\ &= \frac{1}{2}\int \mathrm{d}^3x N\sqrt{\det h}\left(D^c\left(h^{de}\left(D_d\left(\delta^a_{(c}\delta^b_{e)}\delta(x,y)\right) + D_c\left(\delta^a_{(d}\delta^b_{e)}\delta(x,y)\right)\right.\right.\right.\\ &\quad \left.\left.\left.- D_e\left(\delta^a_{(d}\delta^b_{c)}\delta(x,y)\right)\right)\right) + D_d\left(h^{de}h^{cf}\left(D_c\left(\delta^a_{(f}\delta^b_{e)}\delta(x,y)\right)\right.\right.\\ &\quad \left.\left.\left.+ D_f\left(\delta^a_{(c}\delta^b_{e)}\delta(x,y)\right) - D_e\left(\delta^a_{(c}\delta^b_{f)}\delta(x,y)\right)\right)\right)\right)\\ &= \frac{1}{2}\int \mathrm{d}^3x\sqrt{\det h}\left(-(D^cN)\left(D^{(b}\left(\delta^{a)}_c\delta(x,y)\right) + h^{ab}D_c\delta(x,y)\right.\right.\\ &\quad \left.- D^{(a}\left(\delta^{b)}_c\delta(x,y)\right)\right) + (D^eN)\left(D^{(a}\left(\delta^{b)}_e\delta(x,y)\right) + D^{(a}\left(\delta^{b)}_e\delta(x,y)\right) - h^{ab}D_e\delta(x,y)\right)\Big)\\ &= \frac{1}{2}\sqrt{\det h}(h^{ab}D_cD^cN - 2D^aD^bN + h^{ab}D_eD^eN)\\ &= \sqrt{\det h}(h^{ab}D_cD^cN - D^aD^bN) .\end{aligned}$$

Combining all variations, we obtain the final equation of motion

$$\begin{aligned}\dot{p}^{ab} &= -\frac{N\sqrt{\det h}}{16\pi G}\left({}^{(3)}R^{ab} - \frac{1}{2}{}^{(3)}Rh^{ab}\right) + \frac{8\pi GN}{\sqrt{\det h}}h^{ab}\left(p^{cd}p_{cd} - \frac{1}{2}(p^c_c)^2\right)\\ &\quad - \frac{32\pi GN}{\sqrt{\det h}}\left(p^{ac}p_c{}^b - \frac{1}{2}p^{ab}p^c_c\right) + \frac{\sqrt{\det h}}{16\pi G}(D^aD^bN - h^{ab}D_cD^cN)\\ &\quad + \sqrt{\det h}D_c(p^{ab}N^c/\sqrt{\det h}) - 2p^{c(a}D_cN^{b)} .\end{aligned} \tag{3.80}$$

So far, we have used only the gravitational part of the action given by S_{EH}. If matter sources are present, they, too, contribute to the action and thus to the canonical constraints.

In particular, the matter Hamiltonian will be added to the Hamiltonian constraint $C_{\rm grav}$, and energy flows of matter to the diffeomorphism constraint $C_a^{\rm grav}$. The canonical derivation presents these contributions in terms of functional derivatives of the matter energy, which allows one to relate the terms to physical components of the stress-energy tensor. We will discuss examples in more detail in Chapter 3.6.

Alternative gravity theories Along similar lines, canonical formulations are performed for alternative theories of gravity, such as higher-curvature actions where curvature invariants obtained from contractions of the Riemann tensor are added to the Einstein–Hilbert action, or theories in higher dimensions. With higher-curvature terms, new degrees of freedom normally arise and the phase space is enlarged. In the space-time picture, this results from higher-derivative terms in the action, with field equations no longer of second order. Then, some of the first-order time derivatives are to be treated as independent variables rather than momenta, and they have momenta of their own. One can see this from the variational equations, which for a Lagrangian $L(q, \dot{q}, \ldots, q^{(m)})$ depending on derivatives up to order m are

$$\sum_{k=0}^{m}(-1)^k \frac{\mathrm{d}^k}{\mathrm{d}t^k}\frac{\partial L}{\partial q^{(k)}} = 0\,. \tag{3.81}$$

By the Ostrogradsky procedure, one reformulates this higher-order equation as a first-order system by introducing the variables $p_{m-1} := \partial L/\partial q^{(m)}$ and recursively $p_{k-1} := \partial L/\partial q^{(k)} - \dot{p}_k$ for $1 \le k \le m-1$. The equation (3.81), when combined with the defining equations for all the p_i, then takes the Hamiltonian form $\dot{p}_0 = \partial L/\partial q$. A discussion of Legendre transformation shows that p_k plays the role of the momentum of $q^{(k)}$. This procedure has been reviewed for instance by Govaerts and Rashid (1994) who have also provided a systematic formulation in the spirit of constrained systems. The higher-derivative Lagrangian is extended by auxiliary degrees of freedom q_k and μ_k, $1 \le k \le m-1$, appearing in the extended Lagrangian

$$\bar{L}(q, \dot{q}, q_k, \dot{q}_k, \mu_k) := L(q, q_k, \dot{q}_{m-1}) + \sum_{k=1}^{m} \mu_k (q_k - \dot{q}_{k-1}) \tag{3.82}$$

where L is the original higher-order Lagrangian with higher derivatives reduced to first-order ones using the auxiliary variables subject to the relations $q_k = \dot{q}_{k-1}$ imposed via constraints. (We identify $q_0 := q$.)

Example 3.11 (Auxiliary variables)
A canonical analysis of (3.82) leads to momenta $p_0 := p = \partial\bar{L}/\partial q$, $p_k = \partial\bar{L}/\partial\dot{q}_k$ of q_k and $\pi_k = \partial\bar{L}/\partial\mu_k$ of μ_k, the latter constrained to vanish as primary constraints. Another set of primary constraints follows from the definition of p_k: $\psi_k := p_{k-1} - \mu_k = 0$ for $1 \le k \le m-1$. (Although $\dot{q}_k$ appears in the action, for $k \le m-2$ this happens only

linearly. Momenta are then not constrained to vanish, but they are fixed by simple algebraic equations; only p_{m-1} is a dynamical momentum.) The primary constraints are second class.

A comprehensive analysis of higher-curvature actions for gravity, leading in particular to higher derivatives, has been performed by Deruelle et al. (2009), with further details by Multamäki et al. (2010). An example for higher dimensional theories is given by brane-world models, as reviewed for instance by Maartens (2004). In this interpretation, the 4-dimensional space-time is a submanifold in a higher, say 5-, dimensional one. A canonical analysis then naturally allows a foliation by two independent parameters: one orthogonal to the brane and one along the time flow within the brane. The number of different tensorial quantities according to the double decomposition then rises rapidly and the analysis, as developed by Kovács and Gergely (2006) and Keresztes and Gergely (2010), is quite involved. Another extension of the Einstein–Hilbert action is supergravity, which introduces fermionic degrees of freedom of the gravitational field. A first canonical analysis was given by Deser et al. (1977). We will not cover supergravity in this book, but present the canonical formulation of fermions in Chapter 3.6.4.

3.3.4 Hypersurface-deformation algebra

We have now determined all the constraints of general relativity, four primary as well as four secondary ones. They must form a first-class algebra because we expect there to be four independent gauge transformations by changing space-time coordinates, exactly the number of secondary constraints on phase-space functions. (The primary constraints only generate changes of N and N^a.) Verifying the first-class nature explicitly, especially for the Poisson bracket of two Hamiltonian constraints, is tedious, though in principle straightforward with the canonical structures at hand. It is possible to perform the explicit calculation in simpler form after a canonical transformation to new variables introduced later.

3.3.4.1 Off-shell algebra of constraints

In addition to the first-class nature, which tells us that Poisson brackets of the constraints vanish on the constraint surface, the specific "off-shell" algebra of the constraints, satisfied by the constraint functions on the whole phase space including the part off the constraint surface, is of interest, too. Its form shows us what kinds of transformation the constraints generate, and how they are related to space-time properties. Knowing that the constraints are first class, we can compute the algebra rather easily by using simpler matter actions and their constraints.

In the presence of matter, the combined Hamiltonian constraint is $C = C_{\mathrm{grav}} + C_{\mathrm{matter}}$. Similarly, the diffeomorphism constraint $C_a = C_a^{\mathrm{grav}} + C_a^{\mathrm{matter}}$ receives matter contributions. It is convenient to exhibit the structure of the system for constraints in smeared form, integrated with respect to the multipliers N and N^a: the smeared Hamiltonian constraint $H[N] = \int \mathrm{d}^3x N(x)C(x)$ and the smeared diffeomorphism constraint $D[N^a] = \int \mathrm{d}^3x N^a(x)C_a(x)$. For two Hamiltonian constraints, we then have

$$\begin{aligned}\{H[N], H[M]\} &= \{H_{\mathrm{grav}}[N] + H_{\mathrm{matter}}[N], H_{\mathrm{grav}}[M] + H_{\mathrm{matter}}[M]\}\\ &= \{H_{\mathrm{grav}}[N], H_{\mathrm{grav}}[M]\} + \{H_{\mathrm{grav}}[N], H_{\mathrm{matter}}[M]\}\\ &\quad + \{H_{\mathrm{matter}}[N], H_{\mathrm{grav}}[M]\} + \{H_{\mathrm{matter}}[N], H_{\mathrm{matter}}[M]\}\,.\end{aligned}$$

Now we use the fact that the gravitational Hamiltonian constraint does not contain terms with spatial derivatives of the momenta. If we choose a matter contribution that does not couple to extrinsic curvature but only to the spatial metric, as it is realized for a (minimally coupled) scalar field or the Maxwell Hamiltonian, no integrations by parts have to be performed in computing the mixed bracket $\{H_{\mathrm{grav}}[N], H_{\mathrm{matter}}[M]\}$. Thus, no derivatives of lapse functions occur, and the combination $\{H_{\mathrm{grav}}[N], H_{\mathrm{matter}}[M]\} + \{H_{\mathrm{matter}}[N], H_{\mathrm{grav}}[M]\}$ of mixed brackets is proportional to $NM - MN = 0$. For matter without curvature couplings, we have

$$\{H[N], H[M]\} = \{H_{\mathrm{grav}}[N], H_{\mathrm{grav}}[M]\} + \{H_{\mathrm{matter}}[N], H_{\mathrm{matter}}[M]\}\,. \tag{3.83}$$

Minimally coupled scalar field Here, the second term can be computed easily if we choose a simple scalar field φ with momentum p_φ, subject to the Hamiltonian constraint

$$H_{\mathrm{scalar}}[N] = \int \mathrm{d}^3x N \left(\frac{1}{2}\frac{p_\varphi^2}{\sqrt{\det h}} - \frac{1}{2}\sqrt{\det h}h^{ab}(\partial_a\varphi)(\partial_b\varphi) + \sqrt{\det h}V(\varphi)\right) \tag{3.84}$$

and the diffeomorphism constraint

$$D_{\mathrm{scalar}}[N^a] = \int \mathrm{d}^3x N^a(\partial_a\varphi)p_\varphi\,. \tag{3.85}$$

Functional derivatives of $H_{\mathrm{scalar}}[N]$ are easy to compute:

$$\begin{aligned}\frac{\delta H_{\mathrm{scalar}}[N]}{\delta\varphi(x)} &= \sqrt{\det h}h^{ab}(\partial_a N)(\partial_b\varphi) + N\partial_a\left(\sqrt{\det h}h^{ab}\partial_b\varphi\right) + N\sqrt{\det h}\frac{\mathrm{d}V}{\mathrm{d}\varphi}\\ \frac{\delta H_{\mathrm{scalar}}[N]}{\delta p_\varphi(x)} &= N\frac{p_\varphi}{\sqrt{\det h}}\end{aligned}$$

which gives

$$\begin{aligned}&\frac{\delta H_{\mathrm{scalar}}[N]}{\delta\varphi}\frac{\delta H_{\mathrm{scalar}}[M]}{\delta p_\varphi} - \frac{\delta H_{\mathrm{scalar}}[M]}{\delta\varphi}\frac{\delta H_{\mathrm{scalar}}[N]}{\delta p_\varphi}\\ &= h^{ab}(\partial_b\varphi)p_\varphi(M\partial_a N - N\partial_a M)\end{aligned}$$

and upon integration

$$\{H_{\mathrm{scalar}}[N], H_{\mathrm{scalar}}[M]\} = D_{\mathrm{scalar}}[h^{ab}(M\partial_b N - N\partial_b M)]\,. \tag{3.86}$$

The constraint contributions for a scalar field form a linear algebra, with two Hamiltonian constraints with lapse function N and M having a Poisson bracket equal to the diffeomorphism constraint with shift vector $h^{ab}(M\partial_b N - N\partial_b M)$.

Similarly, we compute $\delta D_{\rm scalar}[N^a]/\delta\varphi = -\partial_a(N^a p_\varphi)$, $\delta D_{\rm scalar}[N^a]/\delta p_\varphi = N^a\partial\varphi$, and thus

$$\begin{aligned}\{D_{\rm scalar}[N^a], D_{\rm scalar}[M^a]\} &= \int {\rm d}^3x \left(-\partial_a(N^a p_\varphi)M^b\partial_b\varphi + N^a\partial_a\varphi\partial_b(M^b p_\varphi)\right)\\ &= D_{\rm scalar}[N^a\partial_a M^b - M^a\partial_a N^b]\,.\end{aligned} \tag{3.87}$$

Since the diffeomorphism constraint of matter is independent of the metric and its momentum, the matter part of the Poisson bracket (3.87) already shows the diffeomorphism constraint algebra. The Poisson bracket of a Hamiltonian constraint with the diffeomorphism constraint, on the other hand, requires the gravitational part of the diffeomorphism constraint to be taken into account due to the metric dependence of the Hamiltonian constraint; see Exercise 3.6.

General algebra Matter contributions do not form individual constraints; they rather contribute to the total constraints together with the gravitational parts. Poisson brackets between matter contributions nevertheless tell us what the full constraint algebra must look like, using the fact that we expect it to be first class: according to (3.83), for matter without curvature couplings the gravitational and matter Poisson brackets must add up to a combination of full constraints. From (3.86), we know that the matter contributions provide a Poisson bracket given by the matter contribution to the diffeomorphism constraint with a specific shift vector. The full algebra can be first class only if the Poisson bracket of two gravitational contributions to the Hamiltonian constraint provides the gravitational part of the diffeomorphism constraint with exactly the same shift vector. Thus, we must have $\{H[N], H[M]\} = -D[h^{ab}(N\partial_b M - M\partial_b N)]$.

Explicit calculations for the Poisson brackets involving the diffeomorphism constraint, such as (3.87), complete the full constraint algebra

$$\{D[N^b], D[M^a]\} = D[\mathcal{L}_{N^b}M^a] \tag{3.88}$$

$$\{D[N^a], H[N]\} = H[\mathcal{L}_{N^a}N] \tag{3.89}$$

$$\{H[N], H[M]\} = -D[h^{ab}(N\partial_b M - M\partial_b N)]\,. \tag{3.90}$$

In the first two lines, we simply have the expected action of infinitesimal spatial diffeomorphisms, with multipliers on the right-hand side given by Lie derivatives $\mathcal{L}_{N^b}M^a = [N, M]^a$ and $\mathcal{L}_{N^a}N = N^a\partial_a N$. The last line, however, does not have a direct geometrical interpretation and differs from the first two by the appearance of the field h^{ab}, a phase-space function rather than just multipliers. This part of the algebra shows that we are dealing here with so-called structure functions, not phase-space independent structure constants as one would have them in a Lie algebra.

3.3.4.2 Space-time and the constraints

Intuitively, the action of the Hamiltonian constraint is supposed to complete the spatial diffeomorphisms generated by $D[N^a]$ to the full space-time diffeomorphisms realized as gauge transformations in general relativity. This is indeed the case, but only "on-shell": space-time diffeomorphisms are generated by the set of all constraints provided that the constraints are solved. One can see the necessity of the on-shell requirement by looking more closely at how the diffeomorphism constraint generates spatial diffeomorphisms: the Poisson bracket of any phase-space function $g(h_{ab}, p^{ab})$ with the diffeomorphism constraint $D[N^a]$ with phase-space independent N^a is the Lie derivative of g along the vector field N^a, $\int d^3x N^a\{g, C_a\} = (\delta g/\delta h_{cd})\mathcal{L}_{N^a} h_{cd} + (\delta g/\delta p^{cd})\mathcal{L}_{N^a} p^{cd}$. However, if we make the shift vector dependent on phase space variables, as it naturally appears in the Poisson bracket of two Hamiltonian constraints, we should more precisely write $\int d^3x N^a\{g, C_a\} = \{g, D[N^a]\} - \int d^3x\{g, N^a\}C_a$ to produce the correct terms for a Lie derivative. The first term is the transformation generated by the diffeomorphism constraint; and the second term $\int d^3x\{g, N^a\}C_a$ vanishes when the diffeomorphism constraint is satisfied, but not off the constraint surface. For phase-space dependent N^a the constraints generate diffeomorphisms only on-shell, and due to the presence of structure functions in (3.90) the constraint algebra cannot be consistently restricted to phase-space independent N and N^a for which the constraints would generate space-time diffeomorphisms even off-shell. The spatial diffeomorphism algebra of $D[N^a]$, on the other hand, can consistently be restricted to phase-space independent N^a: it forms a subalgebra (3.88) not involving structure functions.

Another consequence of the presence of structure functions is that the constraints do not generate a group. Thus, they cannot completely correspond to space-time diffeomorphisms, which do form a group. For the algebra of constraints, not only space-time and its deformations but also the foliation used to set up the splitting and to define lapse and shift is relevant. The decomposition, on the other hand, makes use of the metric, since we use the normal vector to hypersurfaces when we define N and N^a. A space-time diffeomorphism deforms the hypersurfaces, reorients the normal, and mixes lapse and shift. It is not surprising, then, that the metric also shows up in the algebra of constraints. Many of these and related issues are discussed in detail by Hojman et al. (1976). An algebraic notion appropriate for the constraint algebra with structure functions is that of a Lie algebroid; see Chapter 4.3.2.6.

Hypersurface deformations Space-time diffeomorphisms can be realized only if the foliation used to set up the canonical formulation is allowed to be deformed. This is indeed how the algebraic relationship between two Hamiltonian constraints can be visualized (Fig. 3.5). A single action of $H[N]$ is an infinitesimal deformation of a spatial foliation

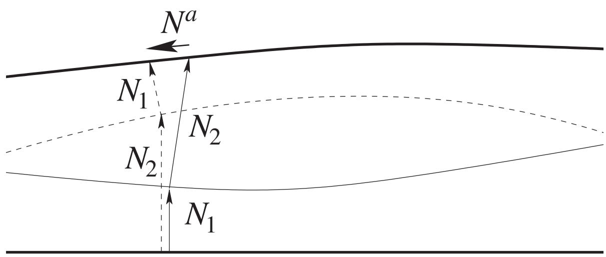

Fig. 3.5 Illustration of the hypersurface deformation algebra. The thin solid line is obtained from the initial one (bottom) by deforming with $N_1(x)$ along the initial normal, the dashed one by deformation with $N_2(x)$. Secondary deformations are then done with respect to the normals of these intermediate surfaces.

along the normal vector field by a position-dependent amount $N(x)$. The Poisson bracket (3.90) corresponds to the difference in doing two such deformations associated with $N(x)$ and $M(x)$ in a row, but in the two different orderings. Since the second deformation is then performed with respect to two different intermediate slicings with their normal vector fields, the result depends on the ordering (unless N and M are constants, in which case the spatial slices are just moved parallel along their normals). Infinitesimally, the difference is a deformation along the final slice as generated by the spatial diffeomorphism on the right-hand side of (3.90).

The hypersurface-deformation algebra illustrates the underlying covariance of the theory. Such an algebraic property of constraints is more complicated to check than covariance of an action formulated in terms of space-time tensors, but it has the same content. (It is even more general, since it can apply to modified space-time structures as well, as they often occur in quantum descriptions of gravity; see Chapter 6.) Just as the gravitational action is determined by the requirements that field equations be covariant and of a certain order in derivatives of the metric as analyzed in detail by Lovelock (1970), the canonical constraints can be reconstructed from the constraint algebra together with an assumption about the derivative order, as shown by Hojman et al. (1976) and Kuchař (1974).

Poincaré algebra For Minkowski space-time, we can use the usual global Cartesian co-ordinates to define spatial slices $t = \text{const.}$, on which we have coordinates x^a. A Poincaré transformation maps these coordinates to new ones t' and $x'^{a'}$, and the new spatial slices $t' = \text{const.}$ represent a hypersurface deformation of the old ones (or in this specific situation, a hypersurface tilt as defined by Kuchař (1976b)). For a pure rotation, only the shift vector is non-vanishing and of the form $N^a = \epsilon^{abc}\gamma_b x_c$. A spatial translation has a constant shift vector $N^a = \delta^a$. For a pure boost, a spatial slice is tilted, represented by a linear lapse function $N = \alpha_a x^a$ vanishing on one plane through the slice with rapidity $\text{artanh}|\alpha|$; see Fig. 3.6. A time translation, finally, has a constant lapse $N = \beta$.

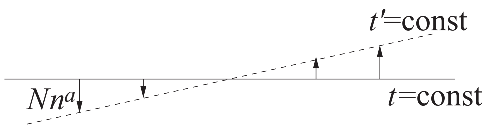

Fig. 3.6 A hypersurface tilt representing a boost by a linear lapse function.

Poincaré transformations must then form a subalgebra of the hypersurface-deformation algebra. Indeed, with lapses and shifts as specified above, one can easily check that all algebraic relations are correct. For instance, for two rotations we have $\{D[N^b], D[M^a]\} = D[\mathcal{L}_{N^b} M^a]$ with a new shift $\mathcal{L}_{N^d} M^a = -\epsilon^{abc}\gamma_b x_c$ with $\gamma_a = \epsilon_{abc}\gamma_1^b\gamma_2^c$ from $N^a = \epsilon^{abc}\gamma_{1b}x_c$, $M^a = \epsilon^{abc}\gamma_{2b}x_c$. A boost and a rotation satisfy $\{H[N], D[N^a]\} = H[N^a\partial_a N]$ with the new lapse $N^a\partial_a N = \epsilon^{abc}\alpha_a\gamma_b x_c$ linear, thus again corresponding to a boost. Finally, for two boosts we have $\{H[N_1], H[N_2]\} = -D[\delta^{ab}(N_1\partial_b N_2 - N_2\partial_b N_1)]$ where the spatial metric is just $h_{ab} = \delta_{ab}$ for a Euclidean slice of Minkowski space. The result is a rotation with linear shift $\delta^{ab}(N_1\partial_b N_2 - N_2\partial_b N_1) = \epsilon^{abc}\gamma_b x_c$ with $\gamma_a = \epsilon_{abc}\alpha_1^b\alpha_2^c$. Similarly, one can see easily that the translation generators are included correctly in the algebra.

3.3.4.3 Space-time gauge and observables

In general relativity we have eight first-class constraints: the four primary constraints $p_\mu \approx 0$, with p_μ canonically conjugate to $N^\mu = (N, N^a)$ now collected in a 4-vector, and the secondary constraints $C_\mu = (C, C_a)$. This suggests the existence of eight independent gauge transformations, even though we expect only four independent infinitesimal space-time coordinate transformations along a vector field ξ^μ. The reason for this apparent doubling of gauge transformations in a canonical framework is the fact that not only coordinates (usually understood as corresponding to the gauge degrees of freedom in general relativity) but also the frame can be changed without affecting observables. Regarding time, for instance, not only the time coordinate t can be changed, which then deforms the foliation of space-time by $t = $ const slices, but also the time-evolution vector field t^a even for a fixed foliation. As we have seen, the foliation does not fix the time-evolution vector field; specifying this vector field thus constitutes extra freedom present in any canonical formulation. A covariant formulation, by contrast, makes use of space-time vector fields and their derivatives along coordinate directions, not separating between space-time splitting (the foliation) and time evolution. In what follows, changing coordinates (and thus the foliation if t is transformed) will be called a *coordinate gauge transformation*, while changing the time-evolution vector field (and thus lapse function and shift vector) will be called a *change of frame*. Such a distinction is useful for a careful treatment of transformations at the Hamiltonian level, while all of them together constitute gauge transformations of the corresponding action.

Space-time gauge For a full discussion of the space-time gauge in canonical gravity, it is useful to employ the extended phase space, where not only h_{ab} and p^{ab} are canonical variables but also N^μ and p_μ. This framework, including its application to gauge transformations and observables, has been developed in particular by Pons et al. (1997). From the perspective of canonical variables, we can see another reason why covariant formulations mix the roles of gauge and frame, distinguished more clearly in canonical versions. If we were to take only the four secondary constraints, they generate the correct transformations of the spatial metric under changes of coordinates. However, the full space-time metric must change, but its time-time and time-space components depend only on lapse and shift and thus have vanishing Poisson brackets with the secondary constraints. In order to generate the correct coordinate transformations of all space-time metric components, the primary constraints are also required. But which combination of primary and secondary constraints generates space-time gauge transformations?

The values of the lapse function and shift vector used in a particular space-time gauge must change when a coordinate transformation is applied. From the canonical perspective, this follows from the fact that lapse and shift enter the equations of motion for the dynamical variables via the constraints. If we consider a generic situation of a fully constrained system with a certain number of constraints denoted collectively as $C[N^A]$ with several independent Lagrange multipliers N^A, any phase-space variable q changes in time by $\dot{q} = \{q, C[N^A]\}$ with respect to the frame specified by a fixed set of N^A. (For general relativity, $N^A = (N(x), N^a(x))$ would be a 4-vector defined on a spatial slice. Counting the spatial dependence, these are infinitely many multipliers.) A coordinate transformation is generated by the same constraints: $q' = q + \delta_{\epsilon^A} q$ with $\delta_{\epsilon^A} q = \{q, C[\epsilon^A]\}$. In general, the transformed q can satisfy the same form of equations of motion,

$$(q')^{\cdot} = \{q', C[N'^A]\}\,, \tag{3.91}$$

only if there is a non-trivial transformation from N^A to $N'^A = N^A + \delta_{\epsilon^B} N^A$. The equations written here, together with the constraint algebra, are sufficient to derive the $\delta_{\epsilon^B} N^A$ and their generator.

We do so by computing both sides of (3.91) separately. On the left, we have

$$\begin{aligned}(q')^{\cdot} &= \dot{q} + (\delta_{\epsilon^A} q)^{\cdot} = \dot{q} + \delta_{\epsilon^A}\dot{q} + \delta_{\dot{\epsilon}^A} q \\ &= \dot{q} + \{\dot{q}, C[\epsilon^A]\} + \{q, C[\dot{\epsilon}^A]\} = \dot{q} + \{\{q, C[N^B]\}, C[\epsilon^A]\} + \{q, C[\dot{\epsilon}^A]\} \\ &= \dot{q} + \{q, \{C[N^B], C[\epsilon^A]\}\} + \{q, C[\dot{\epsilon}^A]\} + \{\{q, C[\epsilon^A]\}, C[N^B]\}\end{aligned}$$

using the Jacobi identity in the last step. On the right-hand side of (3.91), we have

$$\begin{aligned}\{q', C[N^{B\prime}]\} &= \{q + \delta_{\epsilon^A} q, C[N^B + \delta_{\epsilon^C} N^B]\} \\ &= \dot{q} + \{q, C[\delta_{\epsilon^A} N^B]\} + \{\delta_{\epsilon^A} q, C[N^B]\} + O(\epsilon^2) \\ &= \dot{q} + \{q, C[\delta_{\epsilon^A} N^B]\} + \{\{q, C[\epsilon^A]\}, C[N^B]\}\end{aligned}$$

up to terms of second order in the infinitesimal generator ϵ^A. Comparison shows that

$$C[\delta_{\epsilon^A} N^B] = C[\dot{\epsilon}^B] + \{C[N^B], C[\epsilon^A]\} = C[\dot{\epsilon}^B] + C[\langle N, \epsilon\rangle^B] \qquad (3.92)$$

defining the bracket $\langle N, \epsilon\rangle^C := N^A \epsilon^B F^C_{AB}$ with the structure constants (or functions) F^C_{AB} of the constraint algebra $\{C[N^A], C[M^B]\} = C[F^C_{AB} N^A M^B]$. For field theories with smeared constraints $C[N^\mu(x)] = \int \mathrm{d}^3x N^\mu(x) C_\mu(x)$, the definition of the structure constants/functions involves integrations:

$$\{C[N^\mu(x)], C[M^\nu(y)]\} = \int \mathrm{d}^3x \mathrm{d}^3y \mathrm{d}^3z N^\mu(x) N^\nu(y) F^\rho_{\mu\nu}(x, y; z) C_\rho(z) .$$

From (3.92) we read off the change $\delta_{\epsilon^A} N^B = \dot{\epsilon}^B + \langle N, \epsilon\rangle^B$ required for gauge transformations consistent with the equations of motion. Combined with the gauge transformations $\delta_{\epsilon^A} q$ for dynamical fields, a compact expression for the generator of all gauge transformations is $G[\epsilon^A, \dot{\epsilon}^B] = C[\epsilon^A] + p[\dot{\epsilon}^B + \langle N, \epsilon\rangle^B]$ where p_A are again the momenta of N^A, providing the primary constraints now smeared with $\dot{\epsilon}^B + \langle N, \epsilon\rangle^B$. A general gauge transformation, including the Lagrange multipliers, thus depends on ϵ^A as well as $\dot{\epsilon}^A$. These two fields must be considered as independent, since all fields in the canonical formulation are defined only on space, while coordinate time is simply a parameter along the gauge flow. The justification for denoting $\dot{\epsilon}^A$ in this suggestive way comes from a comparison with equations of motion as in the derivation above, or with space-time gauge transformations, for instance as Lie derivatives, in general relativity.

In fact, as one can verify by explicit calculations, the combination

$$G[\epsilon^\mu, \dot{\epsilon}^\mu] = C[\epsilon^\mu] + p\left[\dot{\epsilon}^\mu + \textstyle\int \mathrm{d}^3x \mathrm{d}^3y N^\nu(x) \epsilon^\lambda(y) F^\mu_{\nu\lambda}(x, y; z)\right] , \qquad (3.93)$$

smeared with two independent vector fields ϵ^μ and $\dot{\epsilon}^\mu$ and computed with the secondary constraints and structure functions $F^\mu_{\nu\lambda}(x, y; z)$ of general relativity, is the generator of coordinate changes for all the space-time metric components. The structure functions $F^\mu_{\nu\lambda}(x, y; z)$ can be read off by comparing

$$\{C[\lambda^\mu], C[\kappa^\nu]\} =: C[\textstyle\int \mathrm{d}^3x \mathrm{d}^3y \lambda^\mu(x) \kappa^\nu(y) F^\rho_{\mu\nu}(x, y; z)] \qquad (3.94)$$

with (3.88), (3.89) and (3.90). For instance, $F^0_{a0}(x, y; z) = \delta(x, z)\partial_{y^a}\delta(x, y)$.

In general relativity, the vector field ϵ^μ is defined only on spatial slices, and gives rise to a spatial scalar ϵ^0 parameterizing deformations along the normal of spatial slices, and a spatial vector field ϵ^a parameterizing deformations within spatial slices. For a given foliation with unit normal vector field n^a and spatial vectors s^a_i, $i = 1, 2, 3$, spanning a basis of vector fields tangent to the spatial slices, the components of ϵ^μ provide a space-time vector field

$$\epsilon^a = \epsilon^0 n^a + \epsilon^i s^a_i . \qquad (3.95)$$

The functional (3.93) generates space-time deformations along this vector field. To compare this with a coordinate transformation, we must relate the vector field ϵ^a decomposed in components ϵ^μ normal and tangential to spatial slices, to a vector field with components ξ^μ

along the coordinate lines. Spatial coordinates are not that important in this context, and so we can assume them to change along the directions s_i^a (or assume that we have chosen the s_i^a along spatial coordinate lines in the first place). The time direction, by contrast, matters: in general, it is not along the unit normal vector field n^a but along the time-evolution vector field $t^a = Nn^a + N^a$. Inverting for n^a and inserting in (3.95) provides the new decomposition

$$\epsilon^a = \frac{\epsilon^0}{N} t^a + \left(\epsilon^i - \frac{N^i}{N} \epsilon^0 \right) s_i^a =: \xi^0 t^a + \xi^i s_i^a \,. \tag{3.96}$$

Gauge transformations generated by (3.93) along ϵ^a agree with coordinate transformations $x^\mu \mapsto x^\mu + \xi^\mu$, using the relationship (3.96).

Gauge fixing In order to fix the full gauge, gauge-fixing conditions for the primary as well as secondary constraints must be specified. According to our distinction, this requires conditions to fix the frame (the choice of lapse and shift entering the time-evolution vector field) and the coordinate gauge. It is much easier to discuss frame-fixing, thanks to the simple form of primary constraints.

The primary constraints appear in the total Hamiltonian $H_{\rm total} = H + \lambda^\mu p_\mu$ with unspecified multipliers λ^μ. Fixing the primary constraints means that we choose four additional constraints $\psi_\mu = 0$ such that the combination of primary constraints and the gauge-fixing conditions ψ_μ becomes second class. Second-class constraints can directly be solved without considering any gauge transformation, which in our context means that the primary constraints, and the undetermined multipliers λ^μ with them, will simply drop out of the total Hamiltonian.

With the choice $\psi^\mu := N^\mu - f^\mu$ with certain functions f^μ depending only on h_{ab} and p^{ab} but not on p_μ or N^μ, the primary constraints will clearly be part of a second-class constrained system (p_μ, ψ^ν): $\det\{p_\mu, \psi^\nu\} \neq 0$. Solving the second-class constraints provides the frame-fixed Hamiltonian $H = \int {\rm d}^3x N^\mu C_\mu = \int {\rm d}^3x f^\mu C_\mu$ in which lapse and shift have been specified by f^μ. After frame-fixing, there is a unique generator of the dynamics.

After solving the second-class part of the frame-fixed system, there still remain the secondary constraints C_μ generating the coordinate gauge. They, too, may be fixed by gauge-fixing conditions $\chi_\mu = 0$ such that $\det\{\chi_\mu, C_\nu\} \neq 0$, but doing so explicitly in a globally valid form is much more complicated than for the primary constraints. Moreover, we would want the gauge-fixing conditions to be preserved by the evolution generated by H, imposing the further requirement that

$$\{\chi_\mu, H\} + \frac{\partial \chi_\mu}{\partial t} \approx \int {\rm d}^3x f^\nu \{\chi_\mu, C_\nu\} + \frac{\partial \chi_\mu}{\partial t} = 0 \,.$$

The matrix $\{\chi_\mu, C_\nu\}$ here is required to be invertible for good gauge-fixing conditions, and f^μ cannot vanish identically for a non-trivial dynamics from $H = \int {\rm d}^3x f^\mu C_\mu$. All these conditions can be reconciled only if $\partial\chi_\mu/\partial t \neq 0$ for at least one component μ. In other words, some of the gauge-fixing functions χ_μ of the secondary constraints must be explicitly time-dependent, and thus time in this context must be a combination of phase-space variables: we can solve the gauge-fixing conditions for t (at least locally) to obtain a relationship between t and phase-space variables. Such a time variable is called *internal time*, rather than an external time parameter as used in coordinate representations. Notice that this conclusion, which can be found in Pons et al. (1997), of a required internal-time picture results directly from the fully constrained nature of the system discussed here: only in this case are the matrix $\{\chi_\mu, C_\nu\}$ appearing in the conditions for second-class constraints after gauge-fixing and the one appearing in the dynamical preservation of gauge-fixing conditions identical to each other.

Internal time Coordinate time in general relativity appears as a gauge parameter, and the dynamics is fully constrained. There is no non-trivial Hamiltonian which would take non-zero values on solutions to the constraints and generate any noticeable change with respect to some absolute time. If we decide to forgo gauge parameters and gauge-dependent quantities, time in general relativistic systems must be *relational*: we can notice change not by the motion of objects with respect to an external time parameter, but by the relative change of one object or degree of freedom with respect to another one. This viewpoint has been and is still being developed by Bergmann (1961); Rovelli (1991) and Dittrich (2006, 2007).

The choice of the object whose change is to be described usually depends on the physical questions asked, for instance the growth of the scale factor a in cosmology. Its change, in a gauge-independent manner, must be described with respect to another degree of freedom, such as a matter field φ. The choice of this reference degree of freedom, then called *internal time*, is rather arbitrary. Ideally, one would like a quantity that has a monotonic relationship with coordinate time t in a gauge-dependent formulation. Any event characterized by a fixed value of φ would then correspond to a unique time slice $t = \text{const}$. However, finding such a variable is very difficult, except for some simple models; this constitutes the *problem of time*, reviewed by Kuchař (1992).

The relationship between coordinate-time and internal-time formulations can explicitly be demonstrated for so-called *deparameterizable* systems. They exist in two types: "non-relativistic" and "relativistic". In each case, a system is deparameterizable if its Hamiltonian constraint can be brought to the form

$$C = -f(q, p) + g(q, p)h(p_\varphi) = 0 \tag{3.97}$$

with a function h depending only on the momentum p_φ conjugate to the choice of internal time φ, and two other functions f and g which depend only on canonical variables other than φ and p_φ. For $h(p_\varphi) = p_\varphi$ (or some other linear function), the system is "non-relativistic," for $h(p_\varphi) = p_\varphi^2$ (or some other quadratic polynomial) the system is "relativistic." The distinction here refers to the usual energy dependence of dispersion relations in non-relativistic and relativistic systems, respectively, with p_φ as the momentum of time playing the role of energy. This is a property of the deparameterization, not of the physical model considered. There are "relativistic" as well as "non-relativistic" deparameterizations of models in general relativity, as we will see.

We now describe the "relativistic" version in more detail, but the "non-relativistic" one is very analogous (and slightly simpler). There are then two possible procedures: we can treat the constraint (3.97) as a generator of gauge transformations with coordinate time as the gauge parameter, or we can solve the constraint for p_φ and treat p_φ as a Hamiltonian generating evolution with respect to internal time φ. The latter procedure is reasonable, since the evolution generated for φ, $\mathrm{d}\varphi/\mathrm{d}\varphi = \{\varphi, p_\varphi\} = 1$ makes sure that φ proceeds uniformly. In order to show that p_φ is indeed the Hamiltonian for evolution with respect to φ, we will now evaluate both procedures and verify that they are equivalent.

Let us focus on a phase-space function $x(q, p)$ independent of φ and p_φ. In coordinate time, it changes by

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \{x, C\} = -\{x, f\} + \{x, g\}p_\varphi^2$$

The variable φ changes by

$$\frac{\mathrm{d}\varphi}{\mathrm{d}t} = \{\varphi, C\} = 2gp_\varphi \,.$$

The Hamiltonian $p_\varphi(q, p) = \pm\sqrt{f/g}$ generates evolution with respect to internal time:

$$\frac{\mathrm{d}x}{\mathrm{d}\varphi} = \{x, p_\varphi(q, p)\} = \pm\frac{g^{-1}\{x, f\} - fg^{-2}\{x, g\}}{2\sqrt{f/g}} = -\frac{\mathrm{d}x/\mathrm{d}t}{\mathrm{d}\varphi/\mathrm{d}t} \,.$$

Except for the minus sign, all these equations of motion are consistent with each other according to the chain rule.

Evolving observables The equation of motion $\mathrm{d}x/\mathrm{d}\varphi = \{x, p_\varphi(q, p)\}$ is a differential equation for $x(\varphi)$, possibly coupled to other equations for the remaining canonical variables. If we can find the general solution $x(\varphi; q_0, p_0)$ parameterized by labels q_0, p_0 which one may interpret as the initial values of q and p at some fixed time φ_0, we obtain a gauge-invariant *Dirac observable*: Gauge transformations

$$\{x(\varphi; q_0, p_0), C(q_0, p_0, p_\varphi)\} = -\{x, f\} + \{x, g\}p_\varphi^2 + 2gp_\varphi\frac{\mathrm{d}x}{\mathrm{d}\varphi} = 0$$

of $x(\varphi; q_0, p_0)$, interpreted as a phase-space function via (q_0, p_0) and parameterized by φ, vanish.

Example 3.12 (Free, massless relativistic particle in one dimension)
The energy relation for a free, massless relativistic particle amounts to the constraint $C = -p^2 + p_\varphi^2 = 0$*, where* p_φ *is the energy conjugate to an internal time* φ*. Solving the constraint implies* $p_\varphi = |p|$ *as the Hamiltonian in internal time (assuming the "energy"* p_φ *to be positive). Internal-time equations of motion are* $\mathrm{d}q/\mathrm{d}\varphi = \mathrm{sgn}p$ *and* $\mathrm{d}p/\mathrm{d}\varphi = 0$*. While* p *is constant and already a Dirac observable, for* $q(\varphi)$ *we obtain the solutions*

$q(\varphi; q_0, p_0) = q_0 + \varphi \operatorname{sgn} p_0$. A direct calculation verifies that $\{q(\varphi; q_0, p_0), -p_0^2 + p_\varphi^2\} = -2p_0 + 2p_\varphi \operatorname{sgn} p_0 = 0$.

Equations of motion in coordinate time are $\mathrm{d}q/\mathrm{d}t = -2p$, $\mathrm{d}p/\mathrm{d}t = 0$, $\mathrm{d}\varphi/\mathrm{d}t = 2p_\varphi$ and $\mathrm{d}p_\varphi/\mathrm{d}t = 0$. Clearly, $\mathrm{d}q/\mathrm{d}\varphi = \operatorname{sgn} p = 2p/2|p| = -(\mathrm{d}q/\mathrm{d}t)/(\mathrm{d}\varphi/\mathrm{d}t)$.

3.4 Initial-value problem

For testable predictions, any theory must provide unique solutions in terms of suitable initial data. This corresponds to the practical situation in which one sets up an experiment, that is, specifies an initial configuration and boundary conditions, and then makes observations to be eventually compared with theoretical predictions. A basic mathematical requirement for the equations underlying a theoretical framework is thus that they form a *well-posed initial-value problem*: for any allowed set of initial values a unique solution must exist such that it depends on the initial data in a continuous manner. The first condition ensures predictivity, the latter is added in order to ensure stability in situations in which initial values cannot be arranged to arbitrary precision. For general relativity, analyzing these rather practical questions is closely related to the issue of space-time gauge; see also the review by Friedrich and Rendall (2000).

3.4.1 Hyperbolic systems

Well-posedness is often realized if the underlying differential equations are hyperbolic. We assume a *quasi-linear system* of k differential equations of order m in n variables, of the general form

$$\sum_{\sum_i \alpha_i = m} A_{(\alpha_1,\ldots,\alpha_n)}[u](x)\partial_{x^1}^{\alpha_1}\cdots\partial_{x^n}^{\alpha_n} u = L[u] \tag{3.98}$$

where $u(x) \in \mathbb{R}^k$ collects the unknown functions, $L[\cdot]$ is an arbitrary (possibly non-linear) derivative operator of order $m-1$ and $A_{(\alpha_1,\ldots,\alpha_n)}[u](x)$ are coefficients forming a $k\times k$-matrix depending on x and derivatives $\partial_{x^1}^{\alpha_1}\cdots\partial_{x^n}^{\alpha_n} u$ of lower order $\sum_i \alpha_i < m$. Only the highest-order derivative with $\sum_i \alpha_i = m$, explicitly written in (3.98), thus appears in linear form, but not lower order ones.

For a quasi-linear system, we define the *principal symbol*

$$\sigma(x, u, \xi_a) := \sum_\alpha A_{(\alpha_1,\ldots,\alpha_n)}[u](x)\xi_1^{\alpha_1}\cdots\xi_n^{\alpha_n} \tag{3.99}$$

evaluated in an arbitrary covector ξ_a, whose components are taken to the powers α_i. The principal symbol is thus a $k\times k$-matrix which depends on x, u and the inserted covector. A *characteristic* is a co-vector ξ_a satisfying the *characteristic equation*

$$\det\sigma(x, u, \xi_a) = 0\,. \tag{3.100}$$

A *characteristic surface* is a constant-level surface $\Phi = \text{const}$ for which the differential $\mathrm{d}\Phi$ is a characteristic. In particular, a characteristic surface requires a real-valued characteristic.

Example 3.13 (Wave equation)
For the 2-dimensional Klein–Gordon wave equation $\ddot{\phi} - \phi'' + V(\phi) = 0$ we have the principal symbol $\sigma(\xi) = \xi_t^2 - \xi_x^2$, which is independent of t, x and ϕ. Characteristic surfaces are obtained from functions Φ satisfying $\dot{\Phi}^2 - \Phi'^2 = 0$, solved by $\Phi(t, x) = f(t \pm x)$ with an arbitrary function f. The surfaces $\Phi = \text{const}$ are given by the 2-dimensional light cones $t \pm x = \text{const}$.

In general, characteristic surfaces correspond to wave fronts and determine the local form of linearized plane wave solutions. They are thus relevant for the propagation of initial values.

A quasi-linear system is called *hyperbolic* if km real characteristics exist, counting their multiplicity as solutions of the characteristic equation (3.100). It is called *strictly hyperbolic* if km distinct real characteristics exist. Strictly hyperbolic systems are important because they can be shown to define well-posed initial value problems. For instance, by the preceding example, we know that a single wave equation is strictly hyperbolic. But two coupled wave equations of the same type are only hyperbolic and not strictly hyperbolic because the multiplicity of characteristics is increased. Strictly hyperbolic systems are rather special and do not encompass all types of wave equation found even in simple examples.

A more general version of hyperbolicity, which is implied by strict hyperbolicity, is *symmetric hyperbolicity*. A symmetric hyperbolic system is a set of first-order equations

$$A^0(t, x, u)\partial_t u + A^i(t, x, u)\partial_i u + B(t, x, u) = 0 \tag{3.101}$$

such that A^0 and A^i are symmetric matrices and A^0 is positive definite. Symmetric hyperbolicity is defined only for first-order equations, but this is not a strong restriction, since higher-order equations can be reformulated as first-order ones by introducing auxiliary fields. However, hyperbolicity properties of the resulting system may depend on how the reformulation is performed.

Example 3.14 (Wave equation in symmetric hyperbolic form)
In the wave equation $\ddot{\phi} - \phi'' + V(\phi) = 0$ we introduce $\psi := \dot{\phi}$ and $\sigma := \phi'$ so as to transform it to a set of first-order differential equations. As our independent functions we use ϕ, σ and ψ and impose the additional condition $\dot{\sigma} = \dot{\phi}' = \psi'$ as it follows from the definitions. Then, an equivalent formulation of the wave equation is

$$\partial_t \begin{pmatrix} \phi \\ \sigma \\ \psi \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \partial_x \begin{pmatrix} \phi \\ \sigma \\ \psi \end{pmatrix} + \begin{pmatrix} \psi \\ 0 \\ -V(\phi) \end{pmatrix}$$

which is of symmetric hyperbolic form.

Had we chosen to use $\phi' = \sigma$ as one of the differential equations instead of $\dot{\sigma} = \psi'$, the resulting system would not have been symmetric hyperbolic, although it would certainly *have been equivalent to the same second-order wave equation. This demonstrates how the choice of independent fields matters in the formulation as a symmetric hyperbolic system. In more complicated cases, such as Einstein's equations, exploring the possibilities of bringing a system to symmetric hyperbolic form can be quite involved.*

The example also illustrates the importance of the signature for hyperbolicity: for the differential operator $\ddot{\phi} + \phi''$ *the same procedure would result in a matrix* $A^x = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \end{pmatrix}$ *which is not symmetric. This cannot be repaired by a different choice of independent field because the equation is elliptic, not hyperbolic.*

Symmetric hyperbolic systems are more general than strictly hyperbolic ones; for instance, coupling two wave equations would leave the system symmetric hyperbolic but not strictly hyperbolic. Even the more general symmetric hyperbolic systems are well posed, the proof of which we sketch here for the case of $A^0 = 1$.[7] We will also assume $\partial_i A^i$ and B to be bounded. The key tool in the proof is the *energy functional*

$$E = \int_\Sigma ||u||^2 \mathrm{d}^{n-1}x \tag{3.102}$$

over surfaces $t = \text{const}$. It will allow us to use ordinary differential-equation techniques to analyze partial differential equations.

We assume u to be of compact support and provide an energy estimate, i.e. an upper bound for

$$\frac{\mathrm{d}E}{\mathrm{d}t} = 2\int_\Sigma \langle u, \dot{u}\rangle \mathrm{d}^{n-1}x = -2\int_\Sigma \langle u, A^i \partial_i u + B\rangle \mathrm{d}^{n-1}x\,. \tag{3.103}$$

Using the symmetry of A^i, we have

$$\begin{aligned} \int_\Sigma \langle u, A^i\partial_i u\rangle \mathrm{d}^{n-1}x &= \int_\Sigma \langle A^i u, \partial_i u\rangle \mathrm{d}^{n-1}x \\ &= \int_\Sigma \partial_i\langle A^i u, u\rangle \mathrm{d}^{n-1}x - \int_\Sigma \langle \partial_i(A^i u), u\rangle \mathrm{d}^{n-1}x \\ &= -\int_\Sigma \langle (\partial_i A^i)u, u\rangle \mathrm{d}^{n-1}x - \int_\Sigma \langle A^i\partial_i u, u\rangle \mathrm{d}^{n-1}x \end{aligned}$$

and thus

$$-2\int_\Sigma \langle u, A^i\partial_i u\rangle \mathrm{d}^{n-1}x = \int_\Sigma \langle (\partial_i A^i)u, u\rangle \mathrm{d}^{n-1}x\,.$$

[7] This is a restriction and cannot always be achieved by diagonalization while preserving the symmetric hyperbolic form. One can certainly write $A^0 = R^{-1}DR$ with an orthogonal matrix R and a diagonal one D, since A^0 is symmetric in the symmetric hyperbolic case. Moreover, D cannot have vanishing entries for A^0 to be positive definite; they could thus be transformed to one by rescalings. However, the components of R and D will in general depend on the coordinates and, more importantly in this context, on u, preventing us from absorbing the rotations and rescalings in a redefined dependent vector u without introducing extra terms from derivatives. More general proofs exist to show well-posedness of any strictly hyperbolic system.

Applying the Cauchy–Schwarz inequality to (3.103)

$$
\begin{aligned}
\left|\frac{\mathrm{d}E}{\mathrm{d}t}\right| &= \left|\int_\Sigma \langle(\partial_i A^i)u, u\rangle \mathrm{d}^{n-1}x - 2\int_\Sigma \langle u, B\rangle \mathrm{d}^{n-1}x\right| \\
&\leq \left(\int_\Sigma ||(\partial_i A^i)u - 2B||^2 \mathrm{d}^{n-1}x\right)^{1/2} \left(\int_\Sigma ||u||^2 \mathrm{d}^{n-1}x\right)^{1/2} \\
&\leq ||\partial_i A^i||_{\sup} \int_\Sigma ||u||^2 \mathrm{d}^{n-1}x + 2\left(\int_\Sigma ||B||^2 \mathrm{d}^{n-1}x\right)^{1/2} \left(\int_\Sigma ||u||^2 \mathrm{d}^{n-1}x\right)^{1/2} \\
&\leq C_1 E + C_2\sqrt{E} \qquad (3.104)
\end{aligned}
$$

for some constants C_1 and C_2 obtained as the supremum of $\partial_i A^i$ and the L^2-norm of B, respectively.

This energy estimate can now be used to conclude uniqueness of solutions: assume that we have two solutions u_1 and u_2 for the same initial data and consider $E := \int_\Sigma ||u_1 - u_2||^2 \mathrm{d}^{n-1}x$. Since both solutions have the same initial data, we have that $E_{t=0} = 0$. Moreover, the energy estimate implies that $0 \leq |\mathrm{d}E/\mathrm{d}t| \leq C_1 E + C_2\sqrt{E}$ and thus $(\mathrm{d}E/\mathrm{d}t)_{t=0} = 0$. Taking time derivatives[8] of (3.104) and applying the same reasoning then shows that all time derivatives of E vanish at $E = 0$, and $E(t) = 0$ must be identically satisfied (using analyticity). Thus, $u_1 = u_2$ and any two solutions with the same initial values must be identical. (At this point the positive definiteness of A^0 would be used in a general derivation.)

Similarly, we can derive stability because the energy estimate implies that $\mathrm{d}E/\mathrm{d}t$ must be small if E is small, i.e. nearby solutions deviate from each other only slowly. To demonstrate the existence of solutions for given initial data, we can choose a function basis, approximate it by a finite set and thus map the partial differential equation to a large system of ordinary differential equations. For the latter, there are general results which imply the existence of solutions, and we can use continuity to take the limit back to the infinite basis.

3.4.2 Hyperbolic reductions

In its canonical formulation, Einstein's equation is formulated as a set of first-order equations. The ADM form, however, does not make this system symmetric hyperbolic. For instance, the corresponding A^0 is degenerate: from the point of view of differential equations, N and N^a are independent functions but do not appear with time derivatives in equations of motion. Instead, they can be freely chosen and determine the space-time

[8] An inequality $f_1(x) \leq f_2(x)$ for two differentiable functions does not imply that $f_1'(x) \leq f_2'(x)$ as one can easily see from examples. But if $f_1(0) = f_2(0)$, the inequality does imply that $f_1'(x) \leq f_2'(x)$ in a neighborhood of zero. This is the situation we have here.

frame, that is, the direction of the flow of time used in the differential equations themselves. In the presence of arbitrary functions in the equations, the solution for h_{ab} and p^{ab} starting from given initial data cannot be unique, which would be one of the requirements of well-posedness.

3.4.2.1 Space-time gauge

Especially for numerical relativity — the development of computational codes to solve and evaluate the equations — this constitutes a problem that may be dealt with by including additional equations, such as gauge-fixing conditions. To motivate common choices, it is useful to discuss the relation of N and N^a, or the frame, to the space-time gauge determined by choosing coordinates such that the time evolution vector field is $\partial/\partial t$.

Let us start with an arbitrary coordinate system t', $x'^{a'}$, in which the metric tensor has components $g'^{\mu'\nu'}$. Metric components in canonical form (3.44) are determined by

$$-\frac{1}{N(t,x^a)^2} = g'^{\mu'\nu'}(x')\partial_{\mu'}t\partial_{\nu'}t \tag{3.105}$$

$$\frac{1}{N(t,x^c)^2}N^a(t,x^c) = g'^{\mu'\nu'}(x')\partial_{\mu'}t\partial_{\nu'}x^a \tag{3.106}$$

as they follow from the general tensor transformation law. For a given, specified choice of lapse and shift, these equations may initially not be satisfied, but one may view them as partial differential equations for new coordinates $t(x')$, $x^a(x')$ such that the metric becomes one written in the desired frame: $g^{00} = -1/N^2$ and $g^{0a} = N^a/N^2$ according to (3.44). For instance, (3.105) has the form of a wave equation for t coupled to the remaining equations through x^a. In this way, prescribing N and N^a can be seen as determining coordinate systems, and thus the space-time manifold via the atlas they form.

Solutions to canonical equations depend on initial values for h_{ab} and p^{ab}, but also on N and N^a as gauge conditions (on all of space-time). For given N and N^a, coordinates can be determined in the way described, which implies that the manifold is part of the solution. This is the distinguishing feature of general relativity as a generally covariant framework, in contrast to other field theories such as Maxwell theory whose solutions are functions on a given manifold.

In this sense, general relativity provides dynamical equations not for fields on a given space-time, but for space-time itself. This property is one of immense implications for the fundamental understanding of nature, but it complicates the issue of well-posedness of the field equations. Having to specify certain functions, lapse and shift, on the whole space-time, not just their initial values, is unsuitable for an initial-value problem. Also practically, there is too much freedom in choosing N and N^a to achieve desired properties for space-time representations in numerical codes. For a more tractable set of equations, one has to reformulate the system of equations further by introducing new equations for lapse and shift which determine them through initial values. This can only be achieved by parameterizing the gauge freedom more conveniently, for instance by coupling lapse and shift to *gauge-source functions*: source terms appearing in differential equations for lapse and shift or for coordinates. If a hyperbolic system results in this way, the reformulation is called a *hyperbolic reduction*, discussed by Friedrich (1996).

3.4.2.2 Gauge-source functions and the ADM system

A gauge-source function cannot be formulated in terms of tensors, which would be insensitive to the gauge. Any equation fixing the gauge must be non-covariant so that it can be satisfied only in one coordinate system. A natural choice is to use the non-tensorial Christoffel symbols for this purpose. In particular, we may take the contracted coefficients $\Gamma^{\mu} := g^{\nu\lambda}\Gamma^{\mu}{}_{\nu\lambda}$ and require that they equal a fixed set of functions, $\Gamma^{\mu} = F^{\mu}$. This condition provides differential equations for new coordinates $x'^{\nu'}$, since

$$\nabla_{\mu'}\nabla^{\mu'}x'^{\nu'} = g^{\mu'\lambda'}\nabla_{\mu'}\delta^{\nu'}_{\lambda'} = -g^{\mu'\lambda'}\Gamma^{\rho'}_{\mu'\lambda'}\delta^{\nu'}_{\rho'} = -\Gamma^{\nu'} = -F^{\nu'} . \tag{3.107}$$

(Note that we have to treat $x'^{\nu'}$ as functions rather than a tensor in the first covariant derivative. The second covariant derivative then acts on a tensor provided by the first covariant derivative, which results in the connection coefficients.) Since (3.107) provides standard wave equations known to be well-posed, solutions determine coordinates on space-time from given initial values.

A common choice is $F^{\mu} = 0$, in which case one obtains *harmonic coordinates* satisfying $\nabla_{\mu}\nabla^{\mu}x'^{\nu'} = 0$. These coordinates have several analytical but also numerical advantages as pointed out by Bona and Massó (1988, 1992). The first numerical results obtained by Pretorius (2005) for the form of gravitational waves from merging black holes were extracted using *generalized harmonic coordinates* introduced by Garfinkle (2002), obeying (3.107) with functions $F^{\nu'}$ that are evolved numerically, starting with the harmonic initial condition $F^{\nu'} = 0$. The equations they obey specify the gauge, or the evolution of the frame.

Alternatively, one obtains evolution equations for lapse and shift by prescribing the F^{ν} on all of space-time. From the equation for the Christoffel coefficients in terms of partial derivatives of metric components, applied to the canonical metric, one obtains (see Exercise 3.12)

$$\partial_t N - N^a\partial_a N = N^2(K^a{}_a - n^a\Gamma_a) \tag{3.108}$$

$$\partial_t N^a - N^b\partial_b N^a = N^2(G^a - D^a \log N - h^a_b\Gamma^b) \tag{3.109}$$

where $G^a := h^{bc}G^a{}_{bc}$, $G^a{}_{bc}$ being the spatial connection components as used before. With these equations we have first-order differential equations for lapse and shift in terms of the gauge-source functions Γ^a, which complement the equations of motion (and constraints) we already have for h_{ab} and K_{ab} from the ADM formulation.

In order to see their hyperbolicity, we reformulate all equations as second-order differential equations for N, N^a and h_{ab} for a given Γ^a. The leading differential order in the equation of motion for N is then

$$\begin{aligned}\frac{1}{N^2}(\partial_t - N^a\partial_a)^2 N &= (\partial_t - N^a\partial_a)K^a{}_a + \text{lower-order terms}\\ &= D_a D^a N + \text{lower-order terms}\\ &= h^{ab}\partial_a\partial_b N + \text{ lower-order terms}\end{aligned}$$

where we have used the first-order equation of motion for $K^a{}_a$ and the Hamiltonian constraint to express ${}^{(3)}R$ in terms of lower-order derivatives. Similarly, we obtain

$$\begin{aligned}\frac{1}{N^2}(\partial_t - N^a\partial_a)^2 N^b - h^{cd}\partial_c\partial_d N^a &= \text{lower-order terms}\\ \frac{1}{N^2}(\partial_t - N^a\partial_a)^2 h_{bc} - h^{de}\partial_d\partial_e h_{bc} &= \text{lower-order terms}\end{aligned}$$

with the same differential operators at the highest (second) order. Since lower-order derivatives, even non-linear ones, do not matter for hyperbolicity, they need not be considered here. Then, all fields are subject to the same hyperbolic type of differential equation $g^{ab}\partial_a\partial_b f =$ lower-order terms with the canonical metric g^{ab} from (3.44), coupled to each other only via the lower-order terms.

We have seen that coupling different wave equations of the same type, as now realized by the second-order equations for N, N^a and h_{ab}, does not lead to a strictly hyperbolic system. However, we can formulate all the second-order wave equations as first-order symmetric hyperbolic systems as in Example 3.14. The property of symmetric hyperbolicity is preserved under coupling different equations, which proves that the whole system of our equations is symmetric hyperbolic and thus well-posed. For any given set of gauge-source functions Γ^μ, we obtain a unique solution in terms of initial data. The Γ^μ must be specified on all of space-time, not just on the initial slice. Changing the gauge-source functions implies that we choose different coordinates determined by them as source terms via (3.107). For given initial values, solutions to general relativity are unique up to coordinate changes. The evolution of space-time geometry is uniquely determined.

3.4.2.3 BSSN equations

The ADM system has been used for some time in numerical relativity; see, e.g., some of the results reported by Bona et al. (1995) and Brady et al. (1998). However, it transpires that there are several difficulties regarding a consistent numerical implementation, see, e.g., the discussion by Choptuik (1991) and Frittelli (1997). A useful reformulation of the ADM system was introduced by Baumgarte and Shapiro (1998) and Shibata and Nakamura (1995) by splitting a conformal factor off the spatial metric, and the trace off the extrinsic curvature. That is, instead of h_{ab} and K_{ab}, the variables $h := \det h_{ab}$, $\tilde{h}_{ab} := h^{-1/3}h_{ab}$, $K := h^{ab}K_{ab}$ and $\tilde{K}_{ab} := h^{-1/3}(K_{ab} - \frac{1}{3}Kh_{ab})$ are used in the evolution equations. For instance, we have

$$\dot{K} - N^c D_c K = -N(h^{2/3}\tilde{K}_{ab}\tilde{K}^{ab} + \tfrac{1}{3}K^2) + D^a D_a N \tag{3.110}$$

using (3.80) and the Hamiltonian constraint. (Variables of fractional density weights occur in some of the equations, which is to be taken into account when computing Lie derivatives along the time-evolution vector field.)

So far, only the original ADM system has been reformulated, and the same problems regarding gauge exist. By accompanying the evolution equations for $\log h$, $\tilde{h}_{ab}$, K and $\tilde{K}_{ab}$ with evolution equations for the new gauge-source functions $\tilde{\Gamma}^a := \tilde{h}^{bc}\tilde{\Gamma}^a_{bc}$, computed for the conformal metric $\tilde{h}_{ab}$, a new symmetric hyperbolic system results. Evolution equations for $\tilde{\Gamma}^a = -\partial_b \tilde{h}^{ab}$ (using $\det(\tilde{h}_{ab}) = 1$) are obtained from (3.61) via $\partial_t \tilde{\Gamma}^a = -\partial_b(\partial_t \tilde{h}^{ab})$. In the numerical BSSN codes which have successfully produced black-hole mergers and the resulting gravitational wave-forms, the evolution equations for $\log h$, $\tilde{h}_{ab}$, K, $\tilde{K}_{ab}$ and $\tilde{\Gamma}^a$ are accompanied by modified evolution equations for lapse and shift, rather than (3.108) and (3.109). Campanelli et al. (2006) use a relation similar to (3.108) for $\partial_t N$ and $\partial_t N^a = B^a$ with a new function B^a satisfying $\partial_t B^a = -\frac{3}{4}\partial_t \tilde{\Gamma}^a - \eta B^a$ for some parameter η. In this way, the lapse function can be evolved regularly for a long time. Baker et al. (2006) modify these equations further. At this stage, it is mainly numerical advantages that motivate the choices. For reviews of numerical relativity, see, e.g., Lehner (2001) and Baumgarte and Shapiro (2010).

3.4.3 Slicing conditions

Gauge-source functions such as Γ^μ allow us to prove well-posedness, but they do not always lead easily to intuitive choices that may improve the stability of numerical evolution. Also, directly specifying lapse and shift is often inconvenient. With the simplest choice, $N = 1$ and $N^a = 0$, for instance, one cannot rule out that singularities are approached too quickly. In order to have a good numerical evolution of black holes, valid for long times in regions far away from singularities, one must ensure that the evolved geometries stay away from singularities long enough everywhere on the spatial slices. If we indeed make the choice $N = 1$, $N^a = 0$ in a black-hole space-time of mass M, this would only allow a small amount of proper time of about $\tau = 2M$ for evolution even far away from the horizon. From Eq. (3.110), this choice of frame leads to the equation $\partial K/\partial t \geq \frac{1}{3}K^2$, whose solutions, by the same arguments as used for cosmological models in Chapter 2.4, lead to a coordinate singularity with diverging K after a finite amount of time t. See, also, the detailed analysis by Eardley and Smarr (1979).

In order to find good conditions for lapse and shift, one rather refers to *geometrical slicing conditions* which are formulated more intuitively and can be solved for N and N^a. Several of these concepts have been introduced by Smarr and York (1978) and Garfinkle and Gundlach (1999). The best-known such condition is *maximal slicing*, defined by

$$K = -\frac{8\pi G}{\sqrt{\det h}} p^a{}_a = 0\,, \tag{3.111}$$

the right-hand side in terms of the canonical momentum evolving according to (3.80). Since (3.111) is to hold at all times, the equation

$$\begin{aligned} 0 &= \mathcal{L}_t(p^c{}_c) = h_{ab}\dot{p}^{ab} + p^{ab}\dot{h}_{ab} \\ &= -\frac{\sqrt{\det h}}{8\pi G} D_a D^a N + \mathcal{L}_{N^b} p^a{}_a + \frac{32\pi G N}{\sqrt{\det h}}\left(p^{ab}p_{ab} - \frac{1}{2}(p^c{}_c)^2\right) \end{aligned} \tag{3.112}$$

follows. With the Hamiltonian constraint and (3.111), this provides an elliptic equation

$$D_a D^a N = N^{(3)}R \tag{3.113}$$

for the lapse function which one can use to determine N at fixed times. Thanks to the curvature term, this condition tends to suppress N if ${}^{(3)}R$ becomes large: for such a gauge choice, evolution slows down when singularities are approached. (See also Exercise 3.13.)

The drawback is that slices are stretched and deformed if their evolution is nearly halted at some places but continues unhindered elsewhere. This unwelcome feature can be countered by adapting the shift vector which was free so far. Following Smarr and York (1978), we define the *distortion*

$$\Sigma_{ab} := \frac{1}{2}(\det h)^{1/3}\mathcal{L}_t \frac{h_{ab}}{(\det h)^{1/3}} \tag{3.114}$$

$$= N\left(K_{ab} - \frac{1}{3}h_{ab}K^c{}_c\right) + D_{(a}N_{b)} - \frac{1}{3}h_{ab}D_c N^c \tag{3.115}$$

whose first contribution is the geometric shear, while the rest can be viewed as coordinate shear. Minimizing distortion by varying the functional $D := \int \mathrm{d}^3x\sqrt{\det h}\,\Sigma_{ab}\Sigma^{ab}$ with respect to N^c implies $D^a\Sigma_{ab} = 0$, and thus

$$D_a D^a N_b + \frac{1}{3}D_b D_a N^a + {}^{(3)}R_{ab}N^a = 2D^a\left(N\left(K_{ab} - \tfrac{1}{3}h_{ab}K^c{}_c\right)\right)\,. \tag{3.116}$$

Also here, we have an elliptic equation for the shift vector, whose solution is called *minimal-distortion shift*.

While maximal slicings and minimal distortion shifts are desirable, the elliptic equations they obey are difficult to solve numerically. One often replaces them by driver conditions, as, e.g., by Balakrishna et al. (1996), which are parabolic equations. For maximal slicing, for instance, one can use

$$\partial_t N = \epsilon(D_a D^a N - N^{(3)}R) \tag{3.117}$$

for some parameter ϵ. Solutions to this equation are driven toward the fixed point where the right-hand side vanishes. This also ensures stability of the maximal slicing condition because deviations due to numerical errors cannot grow when the solution is driven back to the fixed point.

3.4.4 Constraints on initial values

By the preceding schemes, evolution from given initial values is formulated in a way useful for numerical procedures. However, evolution equations are not the only equations to be solved: initial values cannot be arbitrary but must obey the constraints. Constraints on spatial fields are not hyperbolic equations but elliptic ones, of the form

$$D^a K_{ab} = 0$$
$$^{(3)}R + (K^a{}_a)^2 - K_{ab}K^{ab} = 0$$

(in vacuum) and thus numerically difficult to deal with. The complexity can, however, be reduced by analytical manipulations.

If we assume maximal slicing, one term already drops out of the Hamiltonian constraint due to $K^a{}_a = 0$. As in the BSSN scheme, we then apply a conformal transformation defining $h_{ab} = \phi^{-4}\tilde{h}_{ab}$, such that $h^{ab} = \phi^4\tilde{h}^{ab}$. Similarly, we rescale the now traceless K_{ab} by $K_{ab} = \phi^{-s}\tilde{K}_{ab}$ for some s to be determined. Here, however, we do not impose conditions on h_{ab} or $\tilde{h}_{ab}$ to fix ϕ; rather, we introduce ϕ as a new degree of freedom obeying an equation to be solved for later on. Starting with any pair (h_{ab}, K_{ab}) with K_{ab} traceless and solving the diffeomorphism constraint, a suitable ϕ can then be found such that the rescaled quantities satisfy the Hamiltonian constraint, as well.

To derive the required equations, we must apply the conformal transformation by ϕ to the curvature components. First, the transformation of the Christoffel symbols can be obtained through the tensor C^c_{ab} in $\tilde{D}_a\omega_b = D_a\omega_b - C^c_{ab}\omega_c$ such that

$$\begin{aligned} C^c_{ab} &= \frac{1}{2}\tilde{h}^{cd}\left(D_a\tilde{h}_{bd} + D_b\tilde{h}_{ad} - D_d\tilde{h}_{ab}\right) \\ &= 2\left(h^c_b D_a \log\phi + h^c_a D_b \log\phi - h_{ab}D^c \log\phi\right) . \end{aligned} \tag{3.118}$$

Then, we have

$$\begin{aligned} \tilde{D}^a\tilde{K}_{ab} &= \phi^{-4}h^{ac}\left(D_c(\phi^s K_{ab}) - C^d_{ca}\phi^s K_{db} - C^d_{cb}\phi^s K_{ad}\right) \\ &= \phi^{-4+s}\left(D^a K_{ab} + (s+2)K_{ab}D^a \log\phi - 2K^a{}_a D_b \log\phi\right) \\ &= \phi^{-4+s}D^a K_{ab} \end{aligned} \tag{3.119}$$

simplified using maximal slicing and setting $s = -2$ at this stage. Finally, transforming the spatial Ricci scalar, the Hamiltonian constraint with maximal slicing (and keeping $s = -2$) becomes

$$^{(3)}\tilde{R} - \tilde{K}_{ab}\tilde{K}^{ab} = -8\phi^{-5}\left(D_aD^a\phi - \frac{1}{8}{}^{(3)}R\phi + \frac{1}{8}\phi^{-7}K_{ab}K^{ab}\right) . \tag{3.120}$$

To find initial values obeying the constraints, we then solve the diffeomorphism constraint $D^a K_{ab} = 0$ for a traceless K_{ab} and some h_{ab} freely specified. According to (3.119), this constraint remains satisfied after the conformal rescaling with $s = -2$. The Hamiltonian constraint, using (3.120) is equivalent to the *Lichnerowicz equation*

$$D_a D^a \phi - \frac{1}{8}{}^{(3)}R\phi + \frac{1}{8}\phi^{-7} K_{ab} K^{ab} = 0 \tag{3.121}$$

for ϕ, using the solution for K_{ab} of the diffeomorphism constraint and the chosen h_{ab}. This equation was introduced by Lichnerowicz (1944) and discussed further, e.g. by York (1971). Then, $\tilde{K}_{ab} = \phi^2 K_{ab}$ and $\tilde{h}_{ab} = \phi^4 h_{ab}$ solve the constraints for maximal slicing.

Example 3.15 (Brill–Lindquist initial data)
Brill and Lindquist (1963) have introduced initial data by starting with a time-symmetric initial surface $\tilde{K}_{ab} = 0$, which obviously solves the vacuum diffeomorphism constraint. Furthermore, h_{ab} is chosen flat such that it only remains to solve the equation $D_a D^a \phi = 0$, the linear Laplace equation. The well-known solution $\phi(r, \vartheta, \varphi) = 1 + M/2r$ in polar coordinates produces the Schwarzschild metric in its conformally flat form.

With a linear equation for ϕ, different solutions, and, in particular, different black holes, can be superposed. The solution

$$\phi(x^a) = 1 + \sum_{i=1}^{N} \frac{\alpha_i}{2\sqrt{(x^a - x_i^a)(x^b - x_i^b)h_{ab}}}\,. \tag{3.122}$$

represents N black holes at initial positions x_i^a. When the distance to all x_i^a is large, the conformal factor behaves as $1 + M/2r$ with the total mass $M = \sum_{i=1}^{N} \alpha_i$. By expanding in the vicinity of a single x_i^a, one obtains the individual masses $M_i = \alpha_i + \sum_{j \neq i} \alpha_i \alpha_j / r_{ij}$ with the distance $r_{ij}^a = \sqrt{(x^a - x_i^a)(x^b - x_i^b)h_{ab}}$ between two centers. Brandt and Brügmann (1997) have extended this method for an application in numerical relativity. Further constructions of general N-body initial data have been provided by Chruściel et al. (2010).

3.5 First-order formulations and Ashtekar variables

The usual formulation of general relativity is of second order: the Einstein–Hilbert action, upon variation by $g_{\mu\nu}$, provides second-order field equations. Alternatively, one can view the system as first order if one initially interprets the connection Γ^a_{bc} as arbitrary, as given by an additional field independent of g_{ab}. In this so-called Palatini formulation,

$$S[g, \Gamma] = \int \mathrm{d}^4 x \sqrt{-\det g}\, R[\Gamma] \tag{3.123}$$

is linear in first-order derivatives of the fields, in particular of Γ, and provides first-order variational equations relating g_{ab} and Γ^a_{bc}.

With Γ^a_{bc} an independent field, new variational equations $\delta S/\delta \Gamma^a_{bc} = 0$ arise and must be satisfied. To compute this variation, assuming for now no Γ^a_{bc}-dependence in the matter terms, we can make use of the earlier result (3.38) and conclude that $\delta R_{ac} = -\nabla_a \delta \Gamma^b_{bc} + \nabla_b \delta \Gamma^b_{ac}$. Inserting this in the variation of (3.123) and integrating by parts shows that the Γ^a_{bc}-variation implies the equation of motion $\nabla_a g_{bc} = 0$.

Example 3.16 (First-order action)
More explicitly, we write

$$R_{ab}[\Gamma] = \partial_c \Gamma^c_{ab} - \partial_b \Gamma^c_{ac} + \Gamma^c_{ab}\Gamma^d_{cd} - \Gamma^c_{ad}\Gamma^d_{bc}$$

and vary the action by Γ^c_{ab}. *We obtain*

$$\begin{aligned} &-\partial_c(\sqrt{-\det g}\, g^{ab}) + g^b_c \partial_d(\sqrt{-\det g}\, g^{ad}) + \sqrt{-\det g}\, g^{ab}\Gamma^d_{cd} \\ &+\sqrt{-\det g}\, g^{de} g^b_c \Gamma^a_{de} - \sqrt{-\det g}\, g^{ad}\Gamma^b_{dc} - \sqrt{-\det g}\, g^{db}\Gamma^a_{dc} = 0 \end{aligned} \tag{3.124}$$

after inserting metric tensors to make all occurrences of Γ^c_{ab} *appear with the same indices. If we take a trace of this equation by equating and summing over c and b, it implies that* $\partial_d(\sqrt{-\det g}\, g^{ad}) = -\sqrt{-\det g}\, g^{de}\Gamma^a_{de}$, *making the second and fourth terms in (3.124) cancel each other. The remaining terms on the left-hand side of (3.124) combine to the covariant derivative* $\nabla_c(\sqrt{-\det g}\, g^{ab})$ *of the tensor density* $\sqrt{-\det g}\, g^{ab}$. *Thus,* $\sqrt{-\det g}\, g^{ab}$ *is covariantly constant, and so is its determinant* $\det(\sqrt{-\det g}\, g^{ab}) = \det g$. *Combining these results,* g^{ab} *and its inverse* g_{ab} *are covariantly constant.*

Solving this equation means that Γ^a_{bc} must be the Christoffel symbol, as the unique one compatible with the metric. In terms of the metric, the connection can thus be solved for, irrespective of initial or boundary conditions. Using the metric-compatible connection in the equations of motion following from varying the metric, which is just $R_{ab}[\Gamma] - \frac{1}{2} g_{ab} R[\Gamma] = 0$, then shows that the same equations of motion result as in the second-order formulation.

The result changes if the connection Γ^a_{bc} does appear in the matter action. Then, matter contributes a source term to the metric compatibility equation and the resulting connection is no longer Christoffel and typically acquires torsion. Such a situation is usually realized when fermionic matter is coupled to gravity, for which the connection must appear in a non-trivial way. (For minimally coupled scalar or Yang–Mills theories, derivative terms are either just partial or exterior derivatives and do not require a connection.) We will come back to this in more detail when discussing matter terms in Chapter 3.6.4.

3.5.1 Tetrad formulation and Holst action

First-order formulations have become particularly important in the Ashtekar framework, which is a canonical formulation of general relativity based on canonical fields given by a triad, instead of the spatial metric, and an independent connection, the Ashtekar connection. From the general viewpoint of constrained systems, this example provides an interesting set of combined first- and second-class constraints.

3.5.1.1 Tetrads and connections

A tetrad provides a way to specify geometries alternative but equivalent to metrics or line elements. Parallel transport or covariant derivatives are then formulated by connections which generalize the Christoffel symbols of the metric formulation. All curvature tensors can be computed in a tetrad formulation as well as in a metric formulation, sometimes even more conveniently so, thanks to the easy use of differential-form notation in tetrad language.

Tetrads We define a *space-time tetrad* as a set of four vector fields e^a_I, labeled by an additional index I taking the values 0, 1, 2, 3, such that they provide an orthonormal basis of the tangent space at each point:

$$g_{ab}e^a_I e^b_J = \eta_{IJ} \tag{3.125}$$

with the Minkowski metric η_{IJ}. Thus, e^a_0 is timelike while e^a_1, e^a_2 and e^a_3 are spacelike. We may view the label I simply as an index to distinguish the vector fields, but a more powerful interpretation is to think of e^a_I as a double-vector field: a vector field on the tangent bundle of space-time taking values in Minkowski space. With a 1-form ω_a, $e^a_I\omega_a$ assigns to every point in space-time a covector in Minkowski space. In this context, one often thinks of space-time as equipped with two independent vectorial structures: its tangent space as well as an independent set of Minkowski spaces, one attached to each point. Minkowski space in this context is called an internal vector space, and $e^a_I\omega_a$ of the above example an internal vector field. A manifold with a copy of the same vector space attached to each point is called a *vector bundle*. (For a precise definition, see the Appendix.) Other examples of vector bundles are used in particle physics, where internal spaces correspond to representation spaces of the gauge groups of fundamental interactions.

Using the Minkowski metric, we can contract the internal indices of a tetrad:

$$\eta^{IJ}e^a_I e_{bJ} = \delta^a_b \tag{3.126}$$

(with $e_{bJ} := g_{bc}e^c_J$) which can easily be verified as a consequence of (3.125) after contraction with the invertible e^b_K. Raising the index b, we have

$$\eta^{IJ}e^a_I e^b_J = g^{ab}\,. \tag{3.127}$$

There are thus orthogonality relationships for both of the possible contractions of a product of tetrad components. So far, we have only lowered the tangent-space index a of a tetrad, using the space-time metric. With the internal Minkowski metric, we may raise the internal index I, too. Doing both operations at once, we have

$$e^I_a = \eta^{IJ}e^b_J g_{ab}\,. \tag{3.128}$$

The orthogonality relations then show that e^I_a is the inverse of e^a_I, just as g^{ab} is the inverse of g_{ab}. We call e^I_a the *co-tetrad*.

In terms of vector spaces — the tangent space $V_p = T_pM$ and the internal Minkowski space M_p at each point p — the tetrad at each point is an isometry $e^a_I(p)\colon M_p \to V_p$, $v^I \mapsto e^a_I(p)v^I$ with inverse map $e^I_a(p)\colon V_p \to M_p$. Without changing the geometry, we can thus replace all tangent-space indices with internal indices, and vice versa, by contractions with the tetrad or co-tetrad: for any tensor field $T^{a_1\cdots a_n}{}_{b_1\cdots b_m}$ we define $T^{I_1\cdots I_n}{}_{J_1\cdots J_m} = e^{I_1}_{a_1}\cdots e^{I_n}_{a_n}e^{b_1}_{J_1}\cdots e^{b_m}_{J_m}T^{a_1\cdots a_n}{}_{b_1\cdots b_m}$.

Equation (3.127) provides another interpretation of the tetrad: it contains all the information found in the metric, since the latter can be reconstructed from it. The tetrad can thus be taken as a fundamental description of geometry, with the metric as a derived concept. However, the tetrad (a tensor without symmetry conditions) has more independent components than the metric. The difference is explained by the fact that we can apply a Lorentz transformation $e^a_I \mapsto \Lambda_I{}^J e^a_J$ to the tetrad without changing the corresponding g^{ab}: by definition of Lorentz transformations, $\eta^{IJ}\Lambda_I{}^K\Lambda_J{}^L e^a_K e^b_L = (\Lambda^T\eta\Lambda)^{KL} e^a_K e^b_L = \eta^{KL} e^a_K e^b_L = g^{ab}$. Lorentz transformations of the tetrad provide new gauge freedom, which we will call internal gauge to distinguish it from the space-time gauge that arises in any theory of space-time geometry.

Connections For vector fields on a curved manifold, parallel transport is not uniquely defined unless a connection is specified. General relativity usually employs the Christoffel connection Γ^a_{bc}, providing the covariant derivative $\nabla_a v^b = \partial_a v^b + \Gamma^b_{ac} v^c$ for vector fields. On a vector bundle, a new type of vector field v^I arises whose parallel transport or covariant derivative must be defined independently of that for space-time vector fields v^a. In order to specify a covariant derivative for internal vector fields, we need *connection 1-forms* $\omega_a{}^I{}_J$, analogous to Γ^b_{ac}, and then define the covariant derivative

$$\mathcal{D}_a v^I = \nabla_a v^I + \omega_a{}^I{}_J v^J \tag{3.129}$$

of a vector field v^I. For internal tensors, each index requires the addition (or subtraction) of a connection term in order to ensure the Leibniz rule. If we have a mixed tangent-space and internal tensor, such as the tetrad, we apply ∇_a using the Christoffel connection for which internal indices do not matter, and a term containing $\pm\omega_a{}^I{}_J$ for every internal index.

Transformation properties of the connection 1-forms can be derived from the behavior of covariant derivatives under local Lorentz transformations by a position-dependent matrix $\Lambda^I{}_J(x)$ changing internal vector fields v^I to $\Lambda^I{}_J v^J$. The space-time covariant derivative does not transform covariantly under local Lorentz transformations, since $\nabla_a(\Lambda^I{}_J v^J) = \Lambda^I{}_J \nabla_a v^J + (\nabla_a \Lambda^I{}_J) v^J$ cannot be written as the Lorentz transformation of an internal vector field. The covariant derivative $\mathcal{D}_a$, on the other hand, has an extra term containing the connection 1-form, which should transform as well if the Lorentz frame is changed. For $\mathcal{D}_a$ to be a covariant derivative, mapping internal tensors to internal tensors, we must have $\mathcal{D}'_a v'^{I'} = \Lambda^{I'}{}_J \mathcal{D}_a v^J$ if $v'^{I'} = \Lambda^{I'}{}_J v^J$ and $\mathcal{D}'_a v'^{I'} = \nabla_a v'^{I'} + \omega'_a{}^{I'}{}_{J'} v^{J'}$ using the transformed connection 1-forms $\omega'_a{}^{I'}{}_{J'}$. Thus,

$$\begin{aligned}
\mathcal{D}'_a(\Lambda^{I'}{}_J v^J) &= \Lambda^{I'}{}_J \nabla_a v^J + (\omega'_a{}^{I'}{}_{J'}\Lambda^{J'}{}_K + \nabla_a \Lambda^{I'}{}_K) v^K \\
&= \Lambda^{I'}{}_J \left(\nabla_a v^J + ((\Lambda^{-1})^J{}_{L'}\omega'_a{}^{L'}{}_{M'}\Lambda^{M'}{}_K + (\Lambda^{-1})^J{}_{L'}\nabla_a\Lambda^{L'}{}_K) v^K\right) \\
&= \Lambda^{I'}{}_J \mathcal{D}_a v^J
\end{aligned}$$

provided that the connection 1-forms transform as

$$\omega_a{}^J{}_K = (\Lambda^{-1})^J{}_{L'}\omega'_a{}^{L'}{}_{M'}\Lambda^{M'}{}_K + (\Lambda^{-1})^J{}_{L'}\nabla_a\Lambda^{L'}{}_K\,. \tag{3.130}$$

Connections 1-forms, just like the Christoffel coefficients, are thus not tensorial but have an inhomogeneous term in their transformation, depending on derivatives of $\Lambda^{L'}{}_K$.

Example 3.17 (Minkowski metric)
The Minkowski metric η_{IJ} is a scalar from the viewpoint of the tangent-space derivative ∇_a, since there are no tangent-space indices. Moreover, it takes the same values everywhere, is spatially constant and thus satisfies $\nabla_a\eta_{IJ} = 0$. But it varies under the covariant derivative $\mathcal{D}_a$, unless the connection 1-forms are such that

$$0 = \mathcal{D}_a\eta_{IJ} = \nabla_a\eta_{IJ} - \omega_a{}^K{}_I\eta_{KJ} - \omega_a{}^K{}_J\eta_{IK} = -\omega_{aJI} - \omega_{aIJ}\,.$$

Connection 1-forms leaving the Minkowski metric invariant must be antisymmetric in their internal indices. We will require this from now on.

With this condition, parallel transport along a curve is a mapping leaving Minkowski space invariant, and amounts to a Lorentz transformation. When contracted with a vector field v^a, the connection 1-forms provide an infinitesimal parallel transport $v^a\omega_a{}^I{}_J$ along the direction v^a. With the required symmetry properties, $v^a\omega_a{}^I{}_J \in \mathrm{so}(1,3)$ indeed takes values in the Lie algebra of SO(1, 3).

We then have a covariant derivative $\mathcal{D}_a$ which preserves the internal metric η_{IJ} as well as, thanks to the same property of ∇_a, the space-time metric g_{ab}. In a tetrad formulation, however, we describe the space-time geometry not by any one of these tensors, but by the tetrad or its inverse, the co-tetrad e_a^I. A reasonable requirement then is that also the co-tetrad must be covariantly constant. This is achieved with the definition

$$\omega_a{}^I{}_J := e^{bI}\nabla_a e_{bJ} \tag{3.131}$$

of the connection 1-forms (which can easily be verified to satisfy the transformation property (3.130).

Example 3.18 (Co-tetrad)
Taking a covariant derivative of the co-tetrad, we have $\mathcal{D}_a e_b^I = \nabla_a e_b^I + \omega_a{}^I{}_J e_b^J$. The connection 1-forms as defined in (3.131) as well as orthogonality of the co-tetrad make the second term equal $\omega_a{}^I{}_J e_b^J = e^{cI}\nabla_a(e_{cJ})e_b^J = -e^{cI}e_{cJ}\nabla_a e_b^J = -\nabla_a e_b^I$. Thus, $\mathcal{D}_a e_b^I = 0$.

The covariant derivative defined with the connection 1-forms (3.131) preserves the tetrad as well as the space-time metric. It must then also preserve the Minkowski metric $\eta^{IJ} = g^{ab}e_a^I e_b^J$ and its inverse η_{IJ}, and so the ω_{aIJ} defined in (3.131) are antisymmetric in I and J, as can easily be verified.

When all indices are made internal, the connection 1-forms $\omega_{IJK} = e_I^a\omega_{aJK} = e_I^a e_J^b\nabla_a e_{bK}$ are called *Ricci rotation coefficients*.

Curvature Starting with the Riemann tensor R_{abcd}, we obtain the internal tensor $R_{IJKL} = R_{abcd}e^a_I e^b_J e^c_K e^d_L = e^a_I e^b_J e^c_K(\nabla_a\nabla_b - \nabla_b\nabla_a)e_{cL}$ by contraction with the tetrad. Using

$$e^c_K\nabla_a\nabla_b e_{cL} = \nabla_a(e^c_K\nabla_b e_{cL}) - (\nabla_a e^c_K)(\nabla_b e_{cL}) = \nabla_a\omega_{bKL} - \eta^{MN}\omega_{aNK}\omega_{bML}$$

we can write the Riemann tensor as

$$\begin{aligned}R_{IJKL} &= e^a_I e^b_J\left(\nabla_a\omega_{bKL} - \nabla_b\omega_{aKL} - \eta^{MN}(\omega_{aNK}\omega_{bML} - \omega_{bNK}\omega_{aML})\right)\\ &= e^a_I\nabla_a\omega_{JKL} - e^a_J\nabla_a\omega_{IKL} - \eta^{MN}(\omega_{INK}\omega_{JML} - \omega_{JNK}\omega_{IML}\\ &\quad + \omega_{INJ}\omega_{MKL} - \omega_{JNI}\omega_{MKL})\,. \end{aligned} \tag{3.132}$$

(Here, we used that e^b_J, while covariantly constant with respect to $\mathcal{D}_a$, is not covariantly constant with respect to ∇_a unless the connection 1-forms all vanish. This provides the extra terms in the last line.) In the last equation, only the Ricci rotation coefficients enter, which are purely internal tensors without space-time indices. They are scalars from the tangent-space point of view, and their space-time covariant derivatives by ∇_a can be simply computed with partial derivatives.

The Riemann tensor then provides the Ricci tensor $R_{IJ} = \eta^{KL}R_{IKJL}$ by contraction. This can easily be seen to agree with the contraction of R_{ab} with tetrad coefficients.

A great advantage of the tetrad formulation is that it lends itself easily to differential-form notation. This often allows faster computations of curvature components or more compact forms of equations of motion. First, the relation for the connection 1-forms, upon antisymmetrization, can be written as $\eta^{IJ}e_{I[a}\omega_{b]KJ} = \nabla_{[a}e_{b]K} = (\mathrm{d}\mathbf{e}_K)_{ab}$ in terms of the exterior derivative of the 1-form $\mathbf{e}_K$ taking values in the internal vector space. The exterior derivative, obtained after antisymmetrization, requires only partial derivatives for its computation, and thus, no knowledge of the Christoffel coefficients is required. Solely in differential forms, we write

$$\mathrm{d}\mathbf{e}_I = \mathbf{e}_J\wedge\boldsymbol{\omega}_I{}^J\,. \tag{3.133}$$

Thanks to the antisymmetry of the connection 1-forms, they can be determined completely from (3.133). (If $\mathrm{d}\mathbf{e}_I$ does not equal $\mathbf{e}_J\wedge\boldsymbol{\omega}_I{}^J$, space-time is said to have *torsion*, with the torsion tensor defined as the difference between these two terms.)

If we introduce the mixed tangent-space/internal-space Riemann tensor R_{abIJ}, we may view it as a 2-form $\mathbf{R}_{IJ}$ thanks to the symmetries of the tensor. Equation (3.132) for the Riemann tensor in terms of connection 1-forms then takes the compact form

$$\mathbf{R}_I{}^J = \mathrm{d}\boldsymbol{\omega}_I{}^J + \boldsymbol{\omega}_I{}^K\wedge\boldsymbol{\omega}_K{}^J\,. \tag{3.134}$$

Equations (3.133) and (3.134) are called the *first and second structure equations*. Analogously, we define the *curvature 2-form* $F_{abI}{}^J = 2\partial_{[a}\omega_{b]I}{}^J + 2\omega_{[a|I|}{}^K\omega_{b]K}{}^J$ of a general connection 1-form $\omega_{aI}{}^J$ not required to obey the first structure equation.

Example 3.19 (Riemann curvature of isotropic models)
A space-time line element of FLRW form, given by (2.1) with (2.2), can be described with a tetrad basis

$$e_\tau^a = -\left(\frac{\partial}{\partial\tau}\right)^a, \qquad e_r^a = \frac{\sqrt{1-kr^2}}{a(\tau)}\left(\frac{\partial}{\partial r}\right)^a$$
$$e_\vartheta^a = \frac{1}{ra(\tau)}\left(\frac{\partial}{\partial\vartheta}\right)^a, \qquad e_\varphi^a = \frac{1}{r\sin\vartheta\, a(\tau)}\left(\frac{\partial}{\partial\varphi}\right)^a .$$

We have chosen the sign of e_τ^a such that the τ-component of the co-tetrad, 1-forms obtained as $\mathbf{e}_I = g_{ab}e_I^b \mathrm{d}x^a = e_{Ia}\mathrm{d}x^a$, is the differential of the time function τ: $\mathbf{e}_\tau = \mathrm{d}\tau$, together with

$$\mathbf{e}_r = \frac{a(\tau)}{\sqrt{1-kr^2}}\mathrm{d}r\,, \qquad \mathbf{e}_\vartheta = a(\tau)r\mathrm{d}\vartheta\,, \qquad \mathbf{e}_\varphi = a(\tau)r\sin\vartheta\mathrm{d}\varphi\,.$$

The first structure equation provides four conditions from the differentials

$$\mathrm{d}\mathbf{e}_\tau = 0 \tag{3.135}$$
$$\mathrm{d}\mathbf{e}_r = \frac{\dot a}{\sqrt{1-kr^2}}\mathrm{d}\tau\wedge\mathrm{d}r \tag{3.136}$$
$$\mathrm{d}\mathbf{e}_\vartheta = \dot a r\mathrm{d}\tau\wedge\mathrm{d}\vartheta + a\mathrm{d}r\wedge\mathrm{d}\vartheta \tag{3.137}$$
$$\mathrm{d}\mathbf{e}_\varphi = \dot a r\sin\vartheta\mathrm{d}\tau\wedge\mathrm{d}\varphi + a\sin\vartheta\mathrm{d}r\wedge\mathrm{d}\varphi + ar\cos\vartheta\mathrm{d}\vartheta\wedge\mathrm{d}\varphi \tag{3.138}$$

from which we are able to read off the Ricci rotation 1-forms $\boldsymbol{\omega}_I{}^J$ uniquely: First, (3.138) can be consistent with (3.133) only if

$$\boldsymbol{\omega}_\varphi{}^\tau = \dot a r\sin\vartheta\mathrm{d}\varphi + \cdots\mathrm{d}\tau$$
$$\boldsymbol{\omega}_\varphi{}^r = \sqrt{1-kr^2}\sin\vartheta\mathrm{d}\varphi + \cdots\mathrm{d}r$$
$$\boldsymbol{\omega}_\varphi{}^\vartheta = \cos\vartheta\mathrm{d}\varphi + \cdots\mathrm{d}\vartheta$$

with coefficients ... as yet undetermined. Equation (3.137) then requires that

$$\boldsymbol{\omega}_\vartheta{}^\tau = \dot a r\mathrm{d}\vartheta + \cdots\mathrm{d}\tau$$
$$\boldsymbol{\omega}_\vartheta{}^r = \sqrt{1-kr^2}\mathrm{d}\vartheta + \cdots\mathrm{d}r$$

and Eq. (3.136)

$$\boldsymbol{\omega}_r{}^\tau = \frac{\dot a}{\sqrt{1-kr^2}}\mathrm{d}r + \cdots\mathrm{d}\tau\,.$$

The first equation (3.135), now solved last, does not provide new 1-forms, but it shows that all unspecified terms "$\cdots$" must vanish. As promised, the Ricci rotation 1-forms have been found uniquely from the first structure equation.

The second structure equation (3.134) allows us to compute the Riemann tensor from the Ricci rotation 1-forms. We have

$$\begin{aligned}
\mathbf{R}_r{}^\tau &= \mathrm{d}\boldsymbol{\omega}_r{}^\tau + \boldsymbol{\omega}_r{}^\vartheta \wedge \boldsymbol{\omega}_\vartheta{}^\tau + \boldsymbol{\omega}_r{}^\varphi \wedge \boldsymbol{\omega}_\varphi{}^\tau \\
&= \frac{\ddot{a}}{\sqrt{1-kr^2}}\mathrm{d}\tau \wedge \mathrm{d}r = \frac{\ddot{a}}{a}\mathbf{e}_\tau \wedge \mathbf{e}_r \\
\mathbf{R}_\vartheta{}^\tau &= \mathrm{d}\boldsymbol{\omega}_\vartheta{}^\tau + \boldsymbol{\omega}_\vartheta{}^r \wedge \boldsymbol{\omega}_r{}^\tau + \boldsymbol{\omega}_\vartheta{}^\varphi \wedge \boldsymbol{\omega}_\varphi{}^\tau \\
&= \ddot{a}r\mathrm{d}\tau \wedge \mathrm{d}\vartheta + \dot{a}\mathrm{d}r \wedge \mathrm{d}\vartheta + \dot{a}\mathrm{d}\vartheta \wedge \mathrm{d}r = \frac{\ddot{a}}{a}\mathbf{e}_\tau \wedge \mathbf{e}_\vartheta \\
\mathbf{R}_\varphi{}^\tau &= \mathrm{d}\boldsymbol{\omega}_\varphi{}^\tau + \boldsymbol{\omega}_\varphi{}^r \wedge \boldsymbol{\omega}_r{}^\tau + \boldsymbol{\omega}_\varphi{}^\vartheta \wedge \boldsymbol{\omega}_\vartheta{}^\tau \\
&= \ddot{a}r\sin\vartheta\mathrm{d}\tau \wedge \mathrm{d}\varphi = \frac{\ddot{a}}{a}\mathbf{e}_\tau \wedge \mathbf{e}_\varphi \\
\mathbf{R}_\vartheta{}^r &= \mathrm{d}\boldsymbol{\omega}_\vartheta{}^r + \boldsymbol{\omega}_\vartheta{}^\tau \wedge \boldsymbol{\omega}_\tau{}^r + \boldsymbol{\omega}_\vartheta{}^\varphi \wedge \boldsymbol{\omega}_\varphi{}^r \\
&= -\frac{kr}{\sqrt{1-kr^2}}\mathrm{d}r \wedge \mathrm{d}\vartheta + \frac{r\dot{a}^2}{\sqrt{1-kr^2}}\mathrm{d}\vartheta \wedge \mathrm{d}r = \frac{\dot{a}^2+k}{a^2}\mathbf{e}_\vartheta \wedge \mathbf{e}_r \\
\mathbf{R}_\vartheta{}^\varphi &= \mathrm{d}\boldsymbol{\omega}_\vartheta{}^\varphi + \boldsymbol{\omega}_\vartheta{}^\tau \wedge \boldsymbol{\omega}_\tau{}^\varphi + \boldsymbol{\omega}_\vartheta{}^r \wedge \boldsymbol{\omega}_r{}^\varphi \\
&= \sin\vartheta\mathrm{d}\vartheta \wedge \mathrm{d}\varphi + \dot{a}^2r^2\sin\vartheta\mathrm{d}\vartheta \wedge \mathrm{d}\varphi - (1-kr^2)\sin\vartheta\mathrm{d}\vartheta \wedge \mathrm{d}\varphi \\
&= \frac{\dot{a}^2+k}{a^2}\mathbf{e}_\vartheta \wedge \mathbf{e}_\varphi \\
\mathbf{R}_\varphi{}^r &= \mathrm{d}\boldsymbol{\omega}_\varphi{}^r + \boldsymbol{\omega}_\varphi{}^\tau \wedge \boldsymbol{\omega}_\tau{}^r + \boldsymbol{\omega}_\varphi{}^\vartheta \wedge \boldsymbol{\omega}_\vartheta{}^r \\
&= -\frac{kr}{\sqrt{1-kr^2}}\sin\vartheta\mathrm{d}r \wedge \mathrm{d}\varphi + \sqrt{1-kr^2}\cos\vartheta\mathrm{d}\vartheta \wedge \mathrm{d}\varphi \\
&\quad + \frac{\dot{a}^2r}{\sqrt{1-kr^2}}\sin\vartheta\mathrm{d}\varphi \wedge \mathrm{d}r + \sqrt{1-kr^2}\cos\vartheta\mathrm{d}\varphi \wedge \mathrm{d}\vartheta = \frac{\dot{a}^2+k}{a^2}\mathbf{e}_\vartheta \wedge \mathbf{e}_r
\end{aligned}$$

Projecting with the tetrad basis, so as to obtain the components $R_{IJK}{}^L = e_I^a e_J^b R_{abK}{}^L$ whose indices we will adorn with a bar to indicate the reference to the tetrad rather than coordinate basis, we have the non-vanishing Riemann-tensor components

$$\begin{aligned}
R_{\bar{r}\bar{\tau}\bar{r}}{}^{\bar{\tau}} = R_{\bar{\vartheta}\bar{\tau}\bar{\vartheta}}{}^{\bar{\tau}} = R_{\bar{\varphi}\bar{\tau}\bar{\varphi}}{}^{\bar{\tau}} &= \frac{\ddot{a}}{a} \\
R_{\bar{\vartheta}\bar{r}\bar{\vartheta}}{}^{\bar{r}} = R_{\bar{\vartheta}\bar{\varphi}\bar{\vartheta}}{}^{\bar{\varphi}} = R_{\bar{\varphi}\bar{r}\bar{\varphi}}{}^{\bar{r}} &= \frac{\dot{a}^2+k}{a^2},
\end{aligned}$$

the non-vanishing Ricci tensor components

$$R_{\bar{r}\bar{r}} = R_{\bar{\vartheta}\bar{\vartheta}} = R_{\bar{\varphi}\bar{\varphi}} = \frac{\ddot{a}}{a} + 2\frac{\dot{a}^2+k}{a^2}, \quad R_{\bar{\tau}\bar{\tau}} = -3\frac{\ddot{a}}{a}$$

and the Ricci scalar

$$R = -R_{\bar{\tau}\bar{\tau}} + R_{\bar{r}\bar{r}} + R_{\bar{\vartheta}\bar{\vartheta}} + R_{\bar{\varphi}\bar{\varphi}} = 6\left(\frac{\ddot{a}}{a} + \frac{\dot{a}^2 + k}{a^2}\right).$$

3.5.1.2 General relativity in tetrad form

In terms of tetrads, a first-order action for general relativity is

$$S[e,\omega] = \frac{1}{16\pi G}\int_M \mathrm{d}^4x\, |e| e^a_I e^b_J F^{IJ}_{ab}(\omega). \tag{3.139}$$

with the space-time tetrad e^a_I, with determinant e^{-1}, and the curvature $F^{IJ}_{ab}(\omega) = 2\partial_{[a}\omega^{IJ}_{b]} + 2\omega^{IK}_{[a}\omega^{LJ}_{b]}\eta_{KL}$ of the so(1,3)-connection 1-forms $\omega_a{}^I{}_J$. The independent fields are e^a_I and $\omega_a{}^I{}_J$. Since $\omega_a{}^I{}_J$ is unrestricted at the level of the action, we distinguish notationally between the Riemann curvature R^{IJ}_{ab} and the curvature F^{IJ}_{ab} of $\omega_a{}^I{}_J$.

Example 3.20 (Einstein's equation from the tetrad formulation)
Before we compute equations of motion, it is useful to rewrite the action using $2ee^{[a}_I e^{b]}_J = \varepsilon^{abcd}\epsilon_{IJKL}e^K_c e^L_d$ *(which can be derived by contracting* $\epsilon_{IJKL}e^I_a e^J_b e^K_c e^L_d = e\varepsilon_{abcd}$ *twice with the tetrad). As an aside, this identity also allows us to write the action in differential-form notation:* $S[e,\omega] = (64\pi G)^{-1}\int_M \epsilon_{IJKL}\mathbf{e}^K\wedge\mathbf{e}^L\wedge\mathbf{F}^{IJ}(\boldsymbol{\omega})$.

The curvature variation is $\delta F^{IJ}_{ab} = 2\mathcal{D}_{[a}\delta\omega^{IJ}_{b]}$, *or* $\delta\mathbf{F}^{IJ} = \mathcal{D}\delta\boldsymbol{\omega}^{IJ}$. *(Only the space-time connection drops out of the antisymmetrized covariant derivative, not the Lorentz connection.) Varying the action (3.139) by* $\omega_a{}^I{}_J$ *and integrating by parts provides*

$$\varepsilon^{abcd}\epsilon_{IJKL}\mathcal{D}_a(e^K_c e^L_d) = 0 \tag{3.140}$$

which is equivalent to the compatibility condition

$$\mathcal{D}_{[a}e^I_{b]} = \partial_{[a}e^I_{b]} + \omega_{[a}{}^I{}_{|J|}e^J_{b]} = 0. \tag{3.141}$$

Varying by e^I_a, *we have*

$$\varepsilon^{abcd}\epsilon_{IJKL}e^I_a F^{JK}_{bc} = 0. \tag{3.142}$$

Equation (3.141) implies that the tetrad is covariantly constant with respect to the covariant derivative defined by ω^{IJ}_a. *This equation is identical to the first structure equation (3.133), and solved for a given tetrad by the compatible connection 1-forms (3.131).*

Computed with the 1-forms preserving the tetrad, $F^{IJ}_{ab} = R^{IJ}_{ab}$ *equals the Riemann tensor. The second equation of motion is then related to the mixed-indices Ricci tensor,* $R^I_a = R^{IJ}_{ab}e^b_J$, *with Ricci scalar* $R = R^I_a e^a_I$. *Using* $e^I_a\varepsilon^{abcd}\epsilon_{IJKL} = \varepsilon^{Ibcd}\epsilon_{IJKL} = 3!ee^b_{[J}e^c_K e^d_{L]}$ *allows us to write (3.142) as* $R^K_b - \frac{1}{2}Re^K_b = 0$. *This is Einstein's equation in vacuum.*

As this example shows, the vacuum action (3.139) depends on the connection via the Riemann tensor when it is computed on-shell, that is, evaluated on solutions to the equations of motion. As we will see in Chapter 3.6.4, however, certain matter couplings may change the form of the connection, and F^{IJ}_{ab} *no longer equals the usual Riemann tensor. For this reason, we will continue to refer to the curvature tensor as* F^{IJ}_{ab}.

We will be led to our next reformulation of the gravitational action by exploring what would have happened if $\epsilon^{IJ}{}_{KL}F_{ab}^{KL}$ had been used in the action instead of just F_{ab}^{IJ}. The connection variation then provides $\epsilon^{abcd}\mathcal{D}_a(e_c^I e_d^J) = 0$ instead of (3.140), also equivalent to the compatibility condition. We are thus still dealing with the connection 1-forms preserving the tetrad, and can write the second equation of motion in terms of the Riemann tensor R_{ab}^{IJ}. That equation now reads

$$0 = \varepsilon^{Iabc}\epsilon_{IJKL}\epsilon^{JK}{}_{MN}R_{ab}^{MN} = -2\varepsilon^{Iabc}R_{abIL}$$

and is identically satisfied by virtue of symmetries of the Riemann tensor. With the extra factor of $\epsilon^{IJ}{}_{KL}$, the solution space is extremely enlarged and contains all pairs of connections and tetrads compatible with each other.

Such an action is not of much physical interest, but it allows us to generalize the Palatini action used so far. We now do not replace the Palatini term with $ee_I^a e_J^b F_{ab}^{KL}(\omega)\epsilon^{IJ}{}_{KL}$, but change it by adding this term (with a free, conventionally chosen pre-factor $-1/2\gamma$):

$$S[e,\omega] = \frac{1}{16\pi G}\int_M \mathrm{d}^4x\ |e|e_I^a e_J^b P^{IJ}{}_{KL}F_{ab}^{KL}(\omega) \tag{3.143}$$

with

$$P^{IJ}{}_{KL} = \delta_K^{[I}\delta_L^{J]} - \frac{1}{2\gamma}\epsilon^{IJ}{}_{KL}\,. \tag{3.144}$$

The connection variation then provides the equation

$$\epsilon^{abcd}\epsilon_{IJKL}P^{KL}{}_{MN}\mathcal{D}_a(e_c^M e_d^N) = 0 \tag{3.145}$$

and still results in the compatibility condition. This equivalence to the previous conditions is most easily seen by noting that the matrix $P^{IJ}{}_{KL}$, interpreted as a mapping from the tensor product of two Minkowski spaces into itself, is invertible, with inverse

$$(P^{-1})_{IJ}{}^{KL} = \frac{\gamma^2}{\gamma^2+1}\left(\delta_I^{[K}\delta_J^{L]} + \frac{1}{2\gamma}\epsilon_{IJ}{}^{KL}\right)\,. \tag{3.146}$$

Varying by the tetrad provides an equation with $\epsilon^{Iabc}R_{abIL}$ added to Einstein's equation. Again, this extra term vanishes by symmetries of the Riemann tensor. Thus, irrespective of the value of γ, we produce the same equations of motion. The action used here is called the Holst action, introduced by Holst (1996), and γ the Barbero–Immirzi parameter whose role was recognized by Barbero G. (1995) and Immirzi (1997). As may be expected from the fact that equations of motion remain unchanged, the extra term in the action can be related to a topological invariant, the Nieh–Yan invariant analyzed in this context by Mercuri (2006) and Date et al. (2009).

In the presence of boundaries, the term $(8\pi G)^{-1}\int_{\partial M}\mathrm{d}^3y\ |e|\ r_a e_I^{[a}e_J^{b]}\omega_b^{IJ}$ with the co-normal r_a to the boundary makes the action functionally differentiable when the tetrad is held fixed at the boundary. Boundary terms and quasilocal energies in the spirit of Brown and York were analyzed in this context of connection and first-order formulations by Lau (1996), for the Palatini action by Ashtekar et al. (2008) and for the Holst action by Corichi and Wilson-Ewing (2010).

3.5.2 Ashtekar–Barbero variables

For a canonical formulation, we foliate space-time. Analogously to introducing the spatial canonical metric h_{ab}, we now consider the new space-time tensor field $\mathcal{E}_I^a = e_I^a + n^a n_I$ with the unit normal n^a to spatial slices and $n_I := e_I^a n_a$. This field satisfies $\mathcal{E}_I^a n_a = \mathcal{E}_I^a n^I = 0$ and thus can be considered a spatial triad. In addition to the usual decomposition of space-time tensors in normal and spatial parts, an extra condition arises to split the internal directions of the tetrad in Minkowski time and space components. The simplest way to do this is by employing a partial gauge fixing of the internal SO(3,1)-transformations, the so-called time gauge. Choosing a timelike internal vector field $n^I = \delta_0^I$, we fix the boost part of internal Lorentz transformations by requiring that $e_0^a = n^I e_I^a = n^a$ is the unit normal to the foliation. The tetrad e_I^a then becomes the local inertial frame of Eulerian observers; the partial gauge fixing is thus natural from the perspective of observables in a canonical formulation arising as those with respect to Eulerian observers. With the partial gauge fixing, internal Minkowski transformations are reduced to spatial rotations by requiring them to fix the chosen n^I. When directly referring only to the spatial rotation part of the group, we will use lower-case internal indices such as $\mathcal{E}_i^a$ for the spatial triad.

In terms of components of the unit normal $n^a = N^{-1}(t^a - N^a)$, we now decompose the action as

$$S[e,\omega] = \frac{1}{16\pi G}\int_M \mathrm{d}^4x\,\sqrt{\det h}\,P^{IJ}{}_{KL}F_{ab}{}^{KL}(\omega)\left(N\mathcal{E}_I^a - 2n_I t^a + 2N^a n_I\right)\mathcal{E}_J^b \tag{3.147}$$

noting the antisymmetry of $P^{IJ}{}_{KL}$. (We have used the determinant $|e| = N\sqrt{\det h}$ of the co-tetrad.)

Given our experience with the ADM action, it is clear that the factor $N\mathcal{E}_I^a - 2n_I t^a + 2N^a n_I$ in (3.147) provides the Hamiltonian constraint by its first term, the diffeomorphism constraint by its last one and the symplectic structure with a derivative along t^a by the middle term. We start with an analysis of the symplectic term to find the new canonical variables: we first introduce the purely spatial tensor

$$P_i^a := \frac{\sqrt{\det h}}{8\pi\gamma G}\mathcal{E}_i^a \tag{3.148}$$

and then write

$$\begin{aligned}
-\gamma\int \mathrm{d}^4x\,n_I t^a P_J^b P^{IJ}{}_{KL}F_{ab}{}^{KL} &= \gamma\int \mathrm{d}^4x\,t^a P_j^b\left(F_{ab}{}^{0j} - \frac{1}{2\gamma}\epsilon^{0j}{}_{KL}F_{ab}{}^{KL}\right)\\
&= \gamma\int \mathrm{d}^4x\,t^a P_j^b\Big(\partial_a\omega_b^{0j} - \partial_b\omega_a^{0j} + 2\omega_{[a}{}^{0k}\omega_{b]k}{}^j\\
&\qquad + \frac{1}{\gamma}\epsilon^j{}_{kl}(\partial_{[a}\omega_{b]}^{kl} + \omega_{[a}{}^{kK}\omega_{b]K}{}^l)\Big)
\end{aligned}$$

where most internal indices have become spatial. (For the signs, notice that $n_I = \eta_{IJ} n^J = \eta_{I0} = -\delta_I^0$ and that we define $\epsilon_{0123} = 1$. The latter choice is irrelevant, since changing the sign can be absorbed by replacing γ with $-\gamma$.) If we integrate by parts the second term, we produce the time derivative $\mathcal{L}_t \omega_b{}^{0j} = t^a \partial_a \omega_b{}^{0j} + \omega_a{}^{0j} \partial_b t^a$ in combination with the first one. Similarly, the second line produces the Lie derivative of $\frac{1}{2\gamma}\epsilon^j{}_{kl}\omega_b^{kl}$. Since these are the only time derivatives appearing in the action, and since both of them are multiplied with γP_j^b, the variable canonically conjugate to P_j^b is

$$A_a^i := \frac{1}{2}\epsilon^i{}_{kl}\omega_a^{kl} + \gamma \omega_a^{0i} \,. \tag{3.149}$$

We can interpret the different space-time connection components entering this variable as follows: $\Gamma_a^i := \frac{1}{2}\epsilon^i{}_{kl}\omega_a^{kl}$ appears in spatial covariant derivatives

$$\mathcal{D}_a v^i = \nabla_a v^i + h_a^b \omega_b{}^i{}_j v^j = \nabla_a v^i - \epsilon^i{}_{jk}\Gamma_a^j v^k \tag{3.150}$$

and is called the *spin connection*. The second term, $K_a^i := \omega_a^{0i}$, can directly be computed from the compatible connection 1-forms (3.131),

$$\mathcal{E}_{ci} K_a^i = -h_a^b \mathcal{E}_{ci} \omega_b^{i0} = -h_a^b \mathcal{E}_{ci} e^{di} \nabla_b e_d^0 = h_a^b h_c^d \nabla_b n_d = K_{ac} \,,$$

(using $e_d^0 = \eta^{0I} g_{dc} e_I^c = -g_{dc} e_0^c = -n_d$) and is then realized as extrinsic curvature contracted with the spatial triad. Thus, our canonical variable is

$$A_a^i = \Gamma_a^i + \gamma K_a^i \,, \tag{3.151}$$

the *Ashtekar–Barbero connection* originally defined by Ashtekar (1987) and Barbero G. (1995).

Example 3.21 (ADM triad variables)
The canonical variables of the ADM formulation using triad rather than metric variables can be obtained from the preceding calculations by the limit $\gamma \to \infty$, which at the level of the action produces the Palatini form. Our canonical variables (A_a^i, P_j^b) found here cannot be directly taken in this limit, but the limit is well defined after the canonical transformation $(A_a^i, P_j^b) \mapsto (\gamma^{-1} A_a^i, \gamma P_j^b)$. For $\gamma \to \infty$, we obtain the canonical pair $(K_a^i, (8\pi G)^{-1} E_j^b)$ with the densitized triad $E_i^a = \sqrt{\det h}\,\mathcal{E}_i^a$ and the extrinsic curvature tensor $K_b^j = \mathcal{E}^{aj} K_{ab}$. In terms of triads, canonical variables do not have to distinguish between extrinsic curvature and momenta, as was required for the p^{ab} of the ADM formulation. This slight simplification is due to the density weight now appearing in metric variables. (Variables of slightly different form had been introduced by Schwinger (1962).) Demonstrating that $(K_a^i, (8\pi G)^{-1} E_j^b) \mapsto (h_{ab}, p^{cd})$ preserves Poisson brackets is the subject of Exercise 3.17.

Continuing with the decomposition of the action, we write the remaining terms in the contribution containing t^a as

$$\int \mathrm{d}^4 x t^a \left(\gamma \omega_a^{0j} \left(\partial_b P_j^b + \omega_{bj}{}^k P_k^b - \frac{1}{\gamma}\epsilon^k{}_{jl}\omega_{b0}{}^l P_k^b\right)\right.$$
$$\left. + \frac{1}{2}\epsilon^j{}_{kl}\omega_a{}^{kl}\left(\partial_b P_j^b - \frac{1}{2}\epsilon_j{}^{nm}\epsilon_{nqp}\omega_b{}^{qp}P_m^b + \gamma\epsilon_{nj}{}^m\omega_b{}^{0n}P_m^b\right)\right)$$
$$= \int \mathrm{d}^4 x \left(\Lambda^j \mathcal{D}_b^{(A)} P_j^b + (1+\gamma^2)\epsilon_{jm}{}^n \omega_t{}^{0j}\omega_b{}^{0m} P_n^b\right)$$

with the covariant derivative $\mathcal{D}_b^{(A)}$ using the Ashtekar–Barbero connection, and introducing $\Lambda^j := \frac{1}{2}\epsilon^j{}_{kl}\omega_t^{kl} + \gamma\omega_t{}^{0j}$. (We have used $\frac{1}{2}\epsilon_{kj}{}^n\epsilon^j{}_{qp}\omega_a{}^{qp}P_n^b = -\omega_{ak}{}^j P_j^b$ and $\epsilon_l{}^{nj}\epsilon^{lkm}\epsilon_{nqp}\omega_{akm}\omega_n{}^{qp}P_j^b = 3!\,\epsilon_l{}^{nj}\delta^l_{[n}\delta^k_q\delta^m_{p]}\omega_{akm}\omega_b{}^{qp}P_j^b = 4\epsilon_l{}^{kj}\ \omega_{ak}{}^m\omega_{bm}{}^l P_j^b$. Notice that Λ^j looks like the t-component of A_a^i, but due to the combination of different $\omega_a{}^{ij}$-components in A_a^i, there is no space-time connection whose pull back is A_a^i.)

The components Λ^j and $\omega_t{}^{0j}$ do not appear with time derivatives in the action; their momenta are thus constrained to vanish, and they provide Lagrange multipliers of secondary constraints. These secondary constraints are the *Gauss constraint*

$$G_j := \mathcal{D}_b^{(A)} P_j^b = 0 \tag{3.152}$$

and the constraint

$$S_j := \epsilon_{jm}{}^n \omega_b{}^{0m} P_n^b = \epsilon_{jm}{}^n K_b^m P_n^b = 0 \tag{3.153}$$

which ensures that $K_{ab} := K_a^i \mathcal{E}_{bi}$ satisfies

$$0 = \epsilon^{ijk} K_b^i P_j^b = K_{ab}\epsilon^{ijk}\mathcal{E}_i^a P_j^b = \frac{1}{8\pi\gamma G} K_{ab}\varepsilon^{abc}\mathcal{E}_c^k\,. \tag{3.154}$$

With the antisymmetric part $K_{[ab]}$ required to vanish, K_{ab} must be a symmetric tensor, compatible with the interpretation as extrinsic curvature. The two constraints (3.152) and (3.153) combined imply also that $\mathcal{D}_b P_j^b = 0$ holds, using only the spin connection. This is consistent with a spin connection compatible with the triad, but the Gauss constraint alone does not uniquely fix the connection. After discussing the remaining constraints, we will be able to determine Γ_a^i uniquely in terms of the triad.

The diffeomorphism and Hamiltonian constraints follow from terms in the action proportional to N^a and N, respectively. These constraints are obtained in terms of purely spatial curvature components $F_{ab}{}^{ij}$, which can be written as

$$F_{ab}^l := \frac{1}{2}\epsilon^l{}_{ij}F_{ab}{}^{ij} = 2\partial_{[a}\Gamma_{b]}^l + \epsilon^l{}_{ij}\epsilon^i{}_{km}\epsilon^{kj}{}_n\Gamma^m_{[a}\Gamma^n_{b]} = 2\partial_{[a}\Gamma^l_{b]} - \epsilon^l{}_{jk}\Gamma^j_{[a}\Gamma^k_{b]} \tag{3.155}$$

using the spin connection. Since the canonical connection is the Asthekar–Barbero one, rather than the spin connection, it is useful to rewrite the curvature using A_a^i, which we denote as

$$\mathcal{F}^l_{ab} = 2\partial_{[a}(\Gamma^l_{b]} + \gamma K^l_{b]}) - \epsilon^l{}_{jk}(\Gamma^j_a + \gamma K^j_a)(\Gamma^k_b + \gamma K^k_b)$$
$$= F^l_{ab} + 2\gamma \mathcal{D}_{[a}K^l_{b]} - \gamma^2 \epsilon^l{}_{jk} K^j_a K^k_b\,. \qquad (3.156)$$

Our contribution to the diffeomorphism constraint then is

$$\begin{aligned} N^a C^{\rm grav}_a &= -\gamma n_I N^a P^b_j P^{IJ}{}_{KL} F_{ab}{}^{KL} \\ &= \gamma N^a P^b_j \left(F_{ab}{}^{0j} - \frac{1}{2\gamma}\epsilon^{0j}{}_{kl} F_{ab}{}^{kl} \right) \\ &= 2\gamma N^a P^b_j \left(\partial_{[a}\omega_{b]}{}^{0j} + \omega_{[a}{}^{0k}\omega_{b]k}{}^j + \frac{1}{2\gamma}\left(2\partial_{[a}\Gamma^j_{b]} + \epsilon^j{}_{kl}\omega_{[a}{}^{kL}\omega_{b]L}{}^l \right) \right) \\ &= N^a P^b_j \left(2\partial_{[a}A^j_{b]} - \gamma \epsilon^j{}_{mk}\Gamma^m_{[a}K^k_{b]} - \frac{1}{2}\epsilon^j{}_{kl}\left(\Gamma^k_{[a}\Gamma^l_{b]} - K^k_{[a}K^l_{b]} \right) \right) \\ &= N^a P^b_j \left(\mathcal{F}^j_{ab} + (1+\gamma^2)\epsilon^j{}_{kl}K^k_a K^l_b \right) \end{aligned} \qquad (3.157)$$

and to the Hamiltonian constraint

$$\begin{aligned} C_{\rm grav} &= -4\pi G\gamma^2 \frac{P^a_i P^b_j}{\sqrt{\det h}} P^{ij}{}_{KL} F_{ab}{}^{KL} \\ &= -4\pi G\gamma^2 \frac{P^a_i P^b_j}{\sqrt{\det h}} \left(F_{ab}{}^{ij} - \frac{1}{2\gamma}\epsilon^{ij}{}_{KL} F_{ab}{}^{KL} \right) \\ &= -4\pi G\gamma^2 \frac{P^a_i P^b_j}{\sqrt{\det h}} \Big(2\partial_{[a}\omega_{b]}{}^{ij} + 2\omega_{[a}{}^{iK}\omega_{b]}{}^{Lj}\eta_{KL} \\ &\qquad + \frac{2}{\gamma}\epsilon^{ij}{}_k \left(\partial_{[a}\omega_{b]}{}^{k0} + \omega_{[a}{}^{kl}\omega_{b]l}{}^0 \right) \Big) \\ &= -4\pi G\gamma^2 \frac{P^a_i P^b_j}{\sqrt{\det h}} \left(F^{ij}_{ab} + 2K^i_{[a}K^j_{b]} - \frac{2}{\gamma}\epsilon^{ij}{}_k \mathcal{D}_{[a}K^k_{b]} \right) \\ &= -4\pi G\gamma^2 \frac{P^a_i P^b_j}{\sqrt{\det h}} \epsilon^{ij}{}_k \Big(\mathcal{F}^k_{ab} + (1+\gamma^2)\epsilon^k{}_{mn}K^m_a K^n_b \\ &\qquad - 2\frac{1+\gamma^2}{\gamma}\epsilon^{ij}{}_k \mathcal{E}^a_i P^b_j \mathcal{D}_{[a}K^k_{b]} \Big)\,. \end{aligned} \qquad (3.158)$$

We have now realized the Hamiltonian as a sum of constraints,

$$H_{\rm grav}[A^i_a, P^b_j] = \int {\rm d}^3x \left(-\Lambda^i G_i - (1+\gamma^2)\omega^{0j}_t S_j + N C_{\rm grav} + N^a C^{\rm grav}_a \right)\,. \qquad (3.159)$$

This expression shows which secondary constraints arise from the non-dynamical variables N, N^a and the connection components Λ^i and ω^{0j}_t. However, there are even more constraints, since we have only the dynamical connection components A^i_a, a fixed combination of ω_a^{ij} and ω_a^{0i}, while ω_a^{ij} and ω_a^{0i} must be allowed to vary independently. A further set of constraints thus comes from the variation by $\Gamma_a^i = \frac{1}{2}\epsilon^i{}_{jk}\omega_a^{jk}$, keeping the dynamical connection A_a^i fixed. In particular, we will have to view K_a^i in the previous terms as the combination $K_a^i = \gamma^{-1}(A_a^i - \Gamma_a^i)$, which is not independent of Γ_a^i. Collecting all terms depending on Γ_a^i in this way, we have another secondary constraint

$$0 = \frac{\delta H_{\rm grav}}{\delta \Gamma_a^i} = -(1+\gamma^2)\epsilon_{ik}{}^l \omega_t^{0k} P_l^a - 2\frac{1+\gamma^2}{\gamma^2} N^{[a} P_j^{b]} \epsilon^j{}_{il}(A_b^l - \Gamma_b^l)$$
$$-\frac{8\pi G(1+\gamma^2)}{\sqrt{\det h}} \left(\epsilon^{jk}{}_i \sqrt{\det h}\, \partial_b \left(P_j^a P_k^b N/\sqrt{\det h}\right) - 2N\Gamma_b^j P_{[i}^a P_{j]}^b\right) .$$

This equation is most useful to analyze after multiplying it with e_{al}:

$$e_{al}\frac{\delta H_{\rm grav}}{\delta \Gamma_a^i} = (1+\gamma^2)\left(\frac{\sqrt{\det h}}{8\pi G\gamma^2}\omega_t{}^{0j}\epsilon_{jil} - \gamma^{-1} e_{al} N^a S_i + \gamma^{-2} N^b \epsilon_{lik}(A_b^k - \Gamma_b^k)\right.$$
$$\left. - \frac{1}{8\pi G}\left(\epsilon_{il}{}^k e_k^b \partial_b N + N\frac{\varepsilon^{abc} e_{al}\partial_b e_{ci}}{\sqrt{\det h}} - N(\delta_{il}\Gamma_b^j \mathcal{E}_j^b - \Gamma_{bl}\mathcal{E}_i^b)\right)\right) .$$

The part symmetric in l and i does not depend on N^a or ω_t^{0j} and is linear in N. Since it must vanish separately, it provides us with the condition

$$\frac{\varepsilon^{abc}}{\sqrt{\det h}}(e_{al}\partial_b e_{ci} + e_{ai}\partial_b e_{cl}) - 2\delta_{il}\Gamma_b^j \mathcal{E}_j^b + \Gamma_{bl}\mathcal{E}_i^b + \Gamma_{bi}\mathcal{E}_l^b = 0\,.$$

The trace of this equation gives $\mathcal{E}_j^b \Gamma_b^j = \frac{1}{2}(\det h)^{-1/2}\varepsilon^{abc} e_{al}\partial_b e_c^l$, and the symmetric part of the condition reduces to an equation for $\Gamma_{b(l}\mathcal{E}_{i)}^b$. From the Gauss constraint together with $S_i = 0$, on the other hand, we have an equation for $\Gamma_{b[l}\mathcal{E}_{i]}^b$. The combination allows us to determine Γ_a^i uniquely in terms of the triad, and it turns out to be the compatible spin connection.

As a consequence, we see that $0 = \delta H_{\rm grav}/\delta\Gamma_a^i$ must at least partially be second class because we are able to solve for some of the non-dynamical variables. Also, the antisymmetric part of the equation can be solved, in this case for the multipliers ω_t^{0k} of S_k; thus, S_k must also be second class. After solving the second-class constraints completely for Γ_a^i, ω_t^{0k} and noting that $S_k = 0$ is implied by the Gauss constraint once the spin connection is known to be compatible with the triad, we are left with

$$H^1_{\rm grav}[A_a^i, P_j^b] = \int {\rm d}^3x \left(-\Lambda^i G_i + N C_{\rm grav} + N^a C_a^{\rm grav}\right) \qquad (3.160)$$

containing only the Gauss, diffeomorphism and Hamiltonian constraints. Moreover, using compatibility of the triad with the spin connection obtained after solving the second-class constraints, we can write

$$C_a^{\mathrm{grav}} = P_j^b \mathcal{F}_{ab}^j + (1+\gamma^2) K_a^k S_k \,, \tag{3.161}$$

$$C_{\mathrm{grav}} = -4\pi G\gamma^2 \epsilon^{ij}{}_k \frac{P_i^a P_j^b}{\sqrt{\det h}} \left(\mathcal{F}_{ab}^k + (1+\gamma^2)\epsilon^k{}_{mn} K_a^m K_b^n\right) - 8\pi G\gamma(1+\gamma^2) \frac{P_i^a \mathcal{D}_a S^i}{\sqrt{\det h}} \,. \tag{3.162}$$

Since $S_i = 0$ once second-class constraints are solved, the diffeomorphism and Hamiltonian constraints can be slightly simplified by dropping their last terms.

The remaining constraints are first class, generating the expected gauge freedom of triad rotations and space-time diffeomorphisms. Since we were able to solve the second-class constraints, H^1_{grav} generates evolution and gauge transformations on the second-class constraint surface; we do not need to compute Dirac brackets.

Example 3.22 (Second-order formulation)
In a second-order formulation, we could use the same action of Holst type but would fix the connection ω_a^{IJ} to be compatible with the triad from the outset. The second-class constraints obtained here by varying the action with respect to connection components do not then arise. Instead of deriving the symmetry of K_{ab} from second-class constraints, it would follow directly from the Gauss constraint.

The results obtained so far can be used to match the number of degrees of freedom used to set up triad and metric formulations. In a triad formulation, the canonical pair (K_b^j, E_i^a) has more degrees of freedom compared to what is realized in a metric formulation with the symmetric tensors (h_{ab}, p^{cd}). The three extra components of the triad as opposed to the spatial metric are removed by the three independent gauge transformations generated by the Gauss constraint. Once the freedom of rotating the triad is removed as gauge, it contains the same information as the spatial metric. Solving the Gauss constraint implies that the tensor K_{ab} constructed from the non-symmetric K_a^i is symmetric. In this way, the degrees of freedom in the different formulations match.

If we allow complex values of the Barbero–Immirzi parameter, complexifying the theory, we notice that the value $\gamma = \pm i$ plays a special role. The constraint algebra in this case changes considerably, since the former second-class constraints, all containing a factor $(1+\gamma^2)$, become identically satisfied. Then, the spin connection remains unrestricted by constraints and can be chosen arbitrarily at least on an initial surface; one would then normally take a second-order viewpoint fixing the spin connection as the compatible one from the outset. (In the covariant analysis of the theory, the freedom of choosing the spin connection follows from the fact that P^{IJ}_{KL} is not invertible for $\gamma = \pm i$, as seen from (3.146).) Furthermore, since K_a^i appear in the first-class constraints only in the combination A_a^i but not separately, the constraints simplify. (This simplification making the constraints polynomial up to a factor of $1/\sqrt{\det h}$ in the Hamiltonian constraint, which could be absorbed using a lapse function of density weight minus one, was the initial motivation for the introduction of complex Ashtekar variables by Ashtekar (1987).) One can solve comparatively simple constraints, and then implement reality conditions. Reality conditions have to require that the densitized triad be real and that the real part of the Ashtekar connection equals the compatible spin connection as determined by the densitized triad. Now, the rather complicated expression for Γ_a^i reappears, not in constraints but in reality conditions. Complications in solving equations cannot be avoided altogether. (If the space-time signature were Euclidean, simplifications would occur for $\gamma = \pm 1$ without the need for reality conditions; see Exercise 3.19.)

3.6 Canonical matter systems

In manifestly covariant formulations of general relativity, the stress-energy tensor is the source of the gravitational field. In a canonical formulation, this space-time tensor appears in a form split up into its components, with energy density as the time-time component, energy flux and momentum density as the time-space components, and the spatial stress tensor (including pressure as one third of its trace) as the purely spatial part. For a given matter field, all these terms can be derived from the matter contributions to the constraints, just as the stress-energy tensor

$$T_{ab} = -\frac{2}{\sqrt{-\det g}} \frac{\delta S_{\mathrm{matter}}}{\delta g^{ab}} \tag{3.163}$$

can be derived by varying the matter action with respect to the metric.

3.6.1 Stress-energy components

A space-time foliation used for the canonical formulation provides a unit normal vector field n^a to the spatial slices, which is the 4-velocity field of Eulerian observers. For matter source terms entering the constraint equations, this is the only vector field that could appear contracted with the stress-energy tensor; the time-evolution vector field only appears when lapse and shift are specified, which is relevant for equations of motion generated by the constraints but not for the constraint equations themselves. Matter contributions to the constraints thus correspond to stress-energy components $T_{ab}n^a n^b$, $T_{ab}n^a h_c^b$ and $T_{ab}h_c^a h_d^b$ projected normally with respect to the slices. These are the components of energy density, energy flux and spatial stress as measured by Eulerian observers.

We will now derive these quantities from a purely canonical perspective. Physically, *energy density* is the amount of energy in a region divided by the size of the region. The energy as seen by a Eulerian observer is the matter contribution to the Hamiltonian constraint, which generates hypersurface deformations along the normal, and the volume is given locally by $\sqrt{\det h}$. Thus, Eulerian observers measure the energy density

$$\rho_{\mathrm{E}} = \frac{H_{\mathrm{matter}}[N]}{N\sqrt{\det h}} \,. \tag{3.164}$$

We have divided by N to ensure that energy is determined with respect to proper time.

Energy current arises when matter is moving with respect to the Eulerian observers. The rest-frame of matter then has a 4-velocity field $t^a/\sqrt{-||t^b||^2}$ with a non-vanishing shift with respect to Eulerian observers. The shift vector enters the diffeomorphism constraint, whose matter contribution gives the energy current

$$J_a^{\mathrm{E}} = \frac{1}{\sqrt{\det h}} \frac{\delta D_{\mathrm{matter}}[N^a]}{\delta N^a} . \tag{3.165}$$

From thermodynamics we have *pressure* defined as the negative change of energy by volume. For a Eulerian observer, the proper energy is obtained as the Hamiltonian by the lapse function, and with volume following from the determinant of the spatial metric, we can write

$$P_{\mathrm{E}} = -\frac{1}{N} \frac{\delta H_{\mathrm{matter}}[N]}{\delta \sqrt{\det h}} . \tag{3.166}$$

In order to rewrite the derivative, we change our variables such that $\det h$ is included as an independent one: $(\det h, \tilde{h}_{ab})$ with $\det \tilde{h} = 1$, thus $h_{ab} = (\det h)^{1/3} \tilde{h}_{ab}$. (This is similar to the treatment of metric components in the BSSN equations. An explicit example to parameterize $\tilde{h}_{ab}$ will be seen in the next chapter in a homogeneous context, given by the Misner variables of Bianchi models.) Keeping $\tilde{h}_{ab}$ fixed in the partial derivative, we have that $\partial h_{ab}/\partial \det h = \frac{1}{3}(\det h)^{-1} h_{ab}$. Then,

$$\frac{\delta}{\delta \sqrt{\det h}} = 2\sqrt{\det h} \frac{\delta}{\delta \det h} = 2\sqrt{\det h} \frac{\partial h_{ab}}{\partial \det h} \frac{\delta}{\delta h_{ab}} = \frac{2}{3\sqrt{\det h}} h_{ab} \frac{\delta}{\delta h_{ab}}$$

and thus

$$P_{\mathrm{E}} = -\frac{2}{3N\sqrt{\det h}} h_{ab} \frac{\delta H_{\mathrm{matter}}[N]}{\delta h_{ab}} . \tag{3.167}$$

In terms of the *spatial stress tensor* S_{E}^{ab}, pressure is defined as $P_{\mathrm{E}} = \frac{1}{3} h_{ab} S_{\mathrm{E}}^{ab}$. Removing the trace from (3.167) thus shows that

$$S_{\mathrm{E}}^{ab} = -\frac{2}{N\sqrt{\det h}} \frac{\delta H_{\mathrm{matter}}[N]}{\delta h_{ab}} \tag{3.168}$$

is an appropriate formula. Of course, the trace does not fix the tensor, but physically it is reasonable that stress components are given by variations of energy by distortions along different directions, as realized by varying metric components.

Example 3.23 (Scalar field)
In the general definitions of canonical stress-energy components, we have used a label "E" to indicate the reference to a special, Eulerian observer. That observer would interpret ρ_{E} as the energy density of matter, and P_{E} as its pressure. These notions are distinct from energy density and pressure of matter as a fluid, which are defined as scalars and without reference to an observer. If the stress-energy tensor of matter has the form $T_{ab} = (\rho + P)u_a u_b + P g_{ab}$ for a unit timelike vector field u_a, ρ is its energy density and P the pressure. Energy density, for instance, is then defined as T_{00} as measured in the rest-frame of the fluid moving

along u^a, a definition that makes ρ invariant. For the constraints of canonical gravity, on the other hand, motion along the unit normal to spatial hypersurfaces is relevant, which cannot always be chosen along matter flows. (An exception is given by isotropic models, in which no non-zero spatial vector can exist. Matter flow in isotropic models thus always has $u^a = n^a$, and Eulerian energy density and pressure agree with the invariantly defined ones. For this reason, we did not include such labels in Chapter 2.)

For a minimally coupled scalar field, the action

$$S_{\text{scalar}} = -\int_M \mathrm{d}^4x\sqrt{-\det g}\,\left(\tfrac{1}{2}g^{ab}(\nabla_a\varphi)(\nabla_b\varphi) + V(\varphi)\right) \tag{3.169}$$

results in the stress-energy tensor

$$T_{ab} = (\nabla_a\varphi)(\nabla_b\varphi) - g_{ab}\left(\tfrac{1}{2}g^{cd}(\nabla_c\varphi)(\nabla_d\varphi) + V(\varphi)\right)\,. \tag{3.170}$$

Provided that $\nabla_a\varphi$ is timelike, we can define $u_a := \nabla_a\phi/\sqrt{-||\nabla_b\varphi||^2}$ and bring the stress-energy tensor in perfect-fluid form with energy density $\rho = -\frac{1}{2}g^{ab}(\nabla_a\varphi)(\nabla_b\varphi) + V(\varphi)$ and pressure $P = -\frac{1}{2}g^{ab}(\nabla_a\varphi)(\nabla_b\varphi) - V(\varphi)$. For a vanishing potential, the scalar obeys the simple equation of state $P = \rho$ of a stiff fluid. This relationship between P and ρ is not satisfied for the quantities measured by general Eulerian observers. (Non-minimal coupling has been analyzed by Madsen (1988).)

We have defined all stress-energy components purely canonically, using the constraints. One can directly verify that $\rho_{\mathrm{E}} = T_{ab}n^an^b$, $J_a^{\mathrm{E}} = -T_{bc}n^bh_a^c$ and $S_{ab}^{\mathrm{E}} = T_{cd}h_a^ch_b^d$ have the correct relationships with projected stress-energy components. To do so, one may start from (3.163), write $\delta S_{\text{matter}} = -\frac{1}{2}\int \mathrm{d}^4x N\sqrt{\det h}T_{ab}\delta g^{ab}$ and compute the components of δg^{ab} in terms of δN, δN^a and δh^{ab} using the canonical metric components introduced by (3.44). For instance, $\delta g^{00} = 2N^{-3}\delta N$ with an index "0" referring to contraction with t^a. Collecting all terms in δS_{matter} multiplied by δN, δN^a and δh^{cd} then relates $\delta S_{\text{matter}}/\delta N$, $\delta S_{\text{matter}}/\delta N^a$ and $\delta S_{\text{matter}}/\delta h^{cd}$, respectively, to the stress-energy components T_{00}, T_{a0} and T_{ab}:

$$\begin{aligned}
T_{00} &= -\frac{N}{\sqrt{\det h}}\left(N\frac{\delta S_{\text{matter}}}{\delta N} + 2N^a\frac{\delta S_{\text{matter}}}{\delta N^a} + 2\frac{N^aN^b}{N^2}\frac{\delta S_{\text{matter}}}{\delta h^{ab}}\right)\\
T_{0a} &= -\frac{N}{\sqrt{\det h}}\left(\frac{\delta S_{\text{matter}}}{\delta N^a} + 2\frac{N^b}{N^2}\frac{\delta S_{\text{matter}}}{\delta h^{ab}}\right)\\
T_{ab} &= -\frac{2}{N\sqrt{\det h}}\frac{\delta S_{\text{matter}}}{\delta h^{ab}}\,.
\end{aligned}$$

Since only metric but no matter variations are involved, we can use $\delta S_{\text{matter}}/\delta g^{ab} = -\delta H_{\text{matter}}/\delta g^{ab}$. Finally, because the components with indices 0 refer to the time-evolution vector field t^a used in (3.44), we must transform to Eulerian observers by expanding the $(0,a)$-components with $n^a = N^{-1}(t^a - N^a)$. In this way we reproduce (3.164), (3.165) and (3.168).

Example 3.24 (Eulerian energy density)
Contracting twice with the normal,

$$T_{ab}n^a n^b = \frac{1}{N^2}(T_{00} - 2T_{0a}N^a + T_{ab}N^a N^b) = \rho_{\mathrm{E}} \tag{3.171}$$

equals the Eulerian energy density (3.164).

Further discussion is given by Kuchař (1976a).

3.6.2 Dust

Dust is a perfect fluid with vanishing pressure. These conditions are idealized, and so dust cannot be considered a fundamental field. Nevertheless, it often provides an interesting matter source for model systems; developing its canonical formulation is thus worthwhile.

As introduced by Brown and Kuchař (1995) (and in a null version by Bičák and Kuchař (1997)), the dynamics of dust can be described using the co-moving coordinates Z^k of its particle constituents as well as proper time T along the flow lines as fields. At an initial time slice, we use spatial coordinates Z^k to provide unique labels for dust-particle world-lines crossing the slice. Proper time T along the world-lines then allows us to complete the initial spatial coordinates to a space-time coordinate system. For a well-defined description, we require $\det(\partial_a Z^k) \neq 0$, excluding intersections of the flow lines (or caustics). The matrix $\partial_a Z^k$ transforms tensor components from arbitrary spatial coordinates to the dust coordinates; it may be viewed as a triad mapping the spatial tangent space to an internal space $\mathbb{R}^3$ equipped with the metric $H^{ij} := h^{ab}(\partial_a Z^i)(\partial_b Z^j)$.

In the dust-coordinate frame, we write the 4-velocity field as $u_a = -\partial_a T + W_i \partial_a Z^i$ with 3-velocity components W_i. The 4-velocity cannot be arbitrary but must satisfy the usual relativistic normalization condition. For this, the fields introduced so far must be constrained by adding a term

$$\mathcal{L}_{\mathrm{dust}} := -\frac{1}{2}\sqrt{-\det g}\,\rho(g^{ab}u_a u_b + 1) \tag{3.172}$$

to the matter Lagrange density, with a multiplier ρ. It turns out that (3.172) is already the complete dust Lagrangian, producing all the required field equations:

- $u^a \partial_a Z^i = 0$ from varying W_i; thus the 4-velocity is indeed along the flow lines of constant Z^i;
- $\partial_a(\sqrt{-\det g}\,\rho u^a) = \sqrt{-\det g}\,\nabla_a(\rho u^a) = 0$ from varying T, thus the current $J^a := \rho u^a$ is conserved;
- $\partial_a(\sqrt{-\det g}\,\rho W_i u^a) = 0$ from varying Z^i, thus also the current $J_i^a := W_i J^a$ is conserved.

A suitable interpretation for ρ is therefore as the energy density, which makes J^a the energy current and J_i^a the momentum current. This interpretation is also consistent with the stress-energy tensor

$$T_{ab} = -\frac{2}{\sqrt{-\det g}}\frac{\delta S}{\delta g^{ab}} = \rho u_a u_b$$

upon using the condition $g^{ab}u_a u_b = -1$ which follows from varying by ρ. We indeed have the stress-energy tensor of a perfect fluid with energy density ρ and vanishing pressure.

With the usual canonical definitions, we decompose the Lagrangian density as

$$\begin{aligned}\mathcal{L}_{\text{dust}} &= \frac{1}{2}\sqrt{\det h}\rho\left(N^{-1}(u_0 - N^a u_a)^2 - N(h^{ab}u_a u_b + 1)\right)\\ &= \frac{1}{2}\sqrt{\det h}\rho\left(N^{-1}(-\dot{T} + W_i\dot{Z}^i - u_a N^a)^2 - N(h^{ab}u_a u_b + 1)\right)\,. \end{aligned} \quad (3.173)$$

Momentum variables are

$$P = \frac{\delta\mathcal{L}_{\text{dust}}}{\delta\dot{T}} = \sqrt{\det h}\frac{\rho}{N}(\dot{T} - W_i\dot{Z}^i + u_a N^a) = -\sqrt{\det h}J^a n_a \quad (3.174)$$

(the Eulerian energy) conjugate to T, and $P_i = \delta\mathcal{L}_{\text{dust}}/\delta\dot{Z}^i = -PW_i$ conjugate to Z^i.

A Legendre transformation leads to the Hamiltonian

$$H_{\text{dust}} = \int \mathrm{d}^3x\left(P\dot{T} + P_i\dot{Z}^i - \mathcal{L}_{\text{dust}}\right) = \int \mathrm{d}^3x\left(NC_{\text{dust}} + N^a C_a^{\text{dust}}\right)$$

with the contribution

$$C_a^{\text{dust}} = -Pu_a = P\partial_a T + P_i\partial_a Z^i \quad (3.175)$$

to the diffeomorphism constraint and

$$C_{\text{dust}} = \frac{1}{2}\frac{P^2}{\rho\sqrt{\det h}} + \frac{1}{2}\rho\sqrt{\det h}\frac{P^2 + h^{ab}C_a^{\text{dust}}C_b^{\text{dust}}}{P^2}$$

to the Hamiltonian constraint. Since $0 = \delta\mathcal{L}_{\text{dust}}/\delta\rho = -N\delta C_{\text{dust}}/\delta\rho$, we have

$$\sqrt{\det h}\rho = \frac{P^2}{\sqrt{P^2 + h^{ab}C_a^{\text{dust}}C_b^{\text{dust}}}}\,.$$

This can be used to simplify C_{dust} as written so far, resulting in

$$C_{\text{dust}} = \sqrt{P^2 + h^{ab}C_a^{\text{dust}}C_b^{\text{dust}}}\,. \quad (3.176)$$

Example 3.25 (Canonical stress-energy of dust)
From the dust contributions to the constraints, we obtain the dust energy density

$$\rho^{\text{E}}_{\text{dust}} = \frac{P}{\sqrt{\det h}}\sqrt{1 + h^{ab}u_a u_b} = -\gamma\rho u^a n_a$$

using (3.175), (3.174) and the relativistic factor $\gamma = \sqrt{1 + h^{ab}u_a u_b} = (1 - h_{ab}V^a V^b)^{-1/2}$ *with the 3-velocity* $V^a = \gamma^{-1}h^{ab}u_b$. *Similarly,*

$$J^{\text{E}}_{a\,\text{dust}} = \frac{Pu_a}{\sqrt{\det h}} = -J^b n_b u_a = \rho u^b n_b u_a\,.$$

In the dust frame with $u_a = n_a$, *these expressions reduce to* $\rho_{\text{dust}} = \rho$ *and* $J_a^{\text{dust}} = -\rho u_a$.

In particular for isotropic models, in which the spatial vector $C_a^{\rm dust} = 0$ must vanish just by the symmetry, we have the simple dust Hamiltonian $C_{\rm dust} = P$ linear in momenta. In this case, the full Hamiltonian constraint $C = C_{\rm grav} + P = 0$ is automatically deparameterized ("non-relativistically") with respect to dust proper time T.

3.6.3 Electromagnetic field

For the electromagnetic field we start with the Maxwell action

$$\begin{aligned} S_{\rm Maxwell} &= -\frac{1}{4}\int {\rm d}^4x\sqrt{-\det g}\,F_{ab}F^{ab} \\ &= -\frac{1}{4}\int {\rm d}^4x\sqrt{-\det g}\,F_{ab}F_{cd}g^{ac}g^{bd} \end{aligned} \tag{3.177}$$

where $F_{ab} = \nabla_a A_b - \nabla_b A_a$ is the field strength of the vector potential A_a. In the last expression in (3.177) it has already been made explicit that the metric is used to raise indices. This is important for the derivation of stress-energy terms, since its components are obtained from derivatives by metric components.

We first transform to the Hamiltonian description. Writing the inverse metric as $g^{ab} = h^{ab} - n^a n^b$ with the spatial metric h_{ab} and the unit normal n^a, we obtain terms such as

$$h_c^b F_{ab} n^a = \frac{1}{N}\left(\dot{A}_b - h_c^b \partial_b\left(A_a t^a\right) - N^a h_c^b F_{ab}\right)\,. \tag{3.178}$$

As earlier, the time derivative $\dot{A}_a = h_a^b \mathcal{L}_t A_b = h_a^b(t^c\partial_c A_b + A_c\partial_b t^c)$ is defined as the Lie derivative along the time-evolution vector field t^a, projected spatially. The momentum conjugate to the spatial A_a, multiplying $\dot{A}_a$ in (3.177), is the densitized vector field

$$E^c = \frac{\delta S_{\rm Maxwell}}{\delta \dot{A}_c} = \sqrt{\det h}\,h^{cd}F_{db}n^b\,. \tag{3.179}$$

One can interpret it as the densitized electric field measured by a Eulerian observer. The time component $t^a A_a$ does not appear as a time derivative and will be a Langrange multiplier for a constraint.

Writing the action as

$$S_{\rm Maxwell} = -\frac{1}{4}\int {\rm d}^4x N\sqrt{\det h}\left(h^{ac}h^{bd}F_{ab}F_{cd} - 2h^{bd}n^a F_{ab}n^c F_{cd}\right) \tag{3.180}$$

and integrating by parts, we obtain

$$\begin{aligned} S_{\rm Maxwell}[A_a, E^a] = \int {\rm d}t \int_{\Sigma_t} {\rm d}^3x\, \Big(& E^a\dot{A}_a + A_d t^d\partial_a E^a - N^c p^a F_{ca} \\ & - N\left(\frac{1}{2\sqrt{\det h}}E^a E^b h_{ab} + \frac{\sqrt{\det h}}{4}F_{ab}F_{cd}h^{ac}h^{bd}\right)\Big) \end{aligned} \tag{3.181}$$

which exhibits all contributions to the constraints. The last two terms, multiplied by N^c and N, respectively, can be recognized as the energy current $S_a = E^b F_{ab} = \varepsilon_{abc}E^b B^c$ and the energy density $\rho = \frac{1}{2}(\det h)^{-1/2}(E^a E_a + B^a B_a)$ of the electromagnetic field. Here, the densitized magnetic field is $B^a = \frac{1}{2}\varepsilon^{abc} F_{bc}$. The contributions add to the diffeomorphism and Hamiltonian constraints when coupled to gravity. The second term in (3.181) is also a constraint, since the time component $A_a t^a$ does not have a non-vanishing momentum. Instead, $A_a t^a$ appears as the multiplier of the Gauss constraint $\partial_a E^a = \nabla_a E^a = 0$ which is just the usual Gauss law of electromagnetism. In smeared form, it generates the gauge transformations of a U(1)-connection: $\{A_a, \int \mathrm{d}^3x \Lambda(x) \nabla_a E^a\} = -\nabla_a \Lambda$.

The total Maxwell Hamiltonian

$$H_{\mathrm{Maxwell}} = \int_{\Sigma_t} \mathrm{d}^3x \Bigg(-A_d t^d \partial_a E^a + N^c E^a F_{ca} + N \left(\frac{1}{2\sqrt{\det h}} E^a E^b h_{ab} + \frac{\sqrt{\det h}}{4} F_{ab} F_{cd} h^{ac} h^{bd} \right) \Bigg) \tag{3.182}$$

can be used to derive stress-energy components. We obtain the energy density

$$\rho_{\mathrm{E}} = \frac{1}{2 \det h} E^a E^b h_{ab} + \frac{1}{4} F_{ab} F_{cd} h^{ac} h^{bd} \tag{3.183}$$

and pressure

$$P_{\mathrm{E}} = \frac{1}{3} \left(\frac{E^a E^b h_{ab}}{2 \det h} + \frac{1}{4} F_{ab} F^{ab} \right) . \tag{3.184}$$

(For a short derivation, one may use the variables $(\det h, \tilde{h}_{ab})$ and take the derivative in (3.167) only by $\det h$.)

The results can be combined to the simple equation of state

$$w = \frac{P_{\mathrm{E}}}{\rho_{\mathrm{E}}} = \frac{1}{3} \tag{3.185}$$

which is always realized for a conformally invariant theory such as electromagnetism. (Such a theory has a vanishing trace of the stress-energy tensor, which requires that $-\rho + 3P = 0$. Since the equation of state in this case follows from the trace of T_{ab}, it is frame-independent.)

3.6.4 Fermions

We have now seen examples for the canonical formulation of matter systems represented by phenomenological fields such as dust, of scalars, and of gauge fields in the simplest case of the U(1)-theory underlying Maxwell's electromagnetism. Non-Abelian gauge theories follow similar lines, as indicated also by the SU(2)-gauge formulation underlying the connection variables for gravity. The remaining class of important matter theories is that of fermions, presented here for the example of Dirac fields.

Dirac fermions are covariantly formulated in terms of 4-component bi-spinors $\Psi = (\psi, \eta)^T$, where ψ and η are 2-spinors transforming in the fundamental representation of $\mathrm{SL}(2, \mathbb{C})$, the complexification of SU(2). In the covariant formulation, we use Dirac's gamma matrices γ_I, defined as satisfying the Clifford algebra relations

$$\gamma_I\gamma_J + \gamma_J\gamma_I = 2\eta_{IJ}\mathbb{I} \tag{3.186}$$

with the Minkowski metric η_{IJ}. For the space-time signature used here, this implies that

$$\gamma_0^2 = -\mathbb{I}\,, \qquad \gamma_j^2 = \mathbb{I}\,, \qquad \gamma_I\gamma_J = -\gamma_J\gamma_I \qquad \text{for} \qquad I \neq J\,. \tag{3.187}$$

An explicit representation satisfying the relations is

$$\gamma^0 = \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix}, \qquad \gamma^j = \begin{pmatrix} 0 & -i\sigma^j \\ i\sigma^j & 0 \end{pmatrix} \tag{3.188}$$

with the Pauli matrices σ^j.

From the four gamma matrices, we construct further useful ones, such as the generators of Lorentz symmetries. Infinitesimal Lorentz transformations $\lambda_I{}^K$, acting on Minkowski vectors v_K, satisfy the relation $\lambda_I{}^K\eta_{KJ} + \eta_{IL}\lambda_J{}^L = 0$, such that λ_{IJ} is antisymmetric. A complete set of antisymmetric matrices is thus sufficient to write arbitrary infinitesimal Lorentz transformations. Using the gamma matrices, one easily finds the antisymmetric combinations $\sigma_{IJ} := \frac{1}{4}[\gamma_I, \gamma_J]$, and the general rule $[[A, B], C] = \{A, \{B, C\}\} - \{B, \{A, C\}\}$ together with the anticommutation relations for gamma matrices easily show $[\gamma_I, \lambda^{JK}\sigma_{JK}] = \lambda_I{}^K\gamma_K$ as required for an infinitesimal Lorentz transformation of a Minkowski vector. The matrices σ_{JK} thus provide representation matrices for infinitesimal Lorentz transformations. Moreover, they satisfy $\sigma_{ij}^\dagger = -\sigma_{ij}$ and $\sigma_{0j}^\dagger = \sigma_{0j}$. On a spinor Ψ, we define the action of the Lorentz group by $\Psi'(x) = S(\Lambda)\Psi(\Lambda^{-1}x)$ where $S(\Lambda) = \exp(\lambda^{IJ}\sigma_{IJ})$ for $\Lambda_I{}^J = \exp(\lambda)_I{}^J$. The matrices $S(\Lambda)$ satisfy the useful relations

$$-\gamma^0 S^\dagger\gamma^0 = S^{-1}\,, \qquad S^{-1}\gamma^I S = \Lambda^I{}_J\gamma^J\,. \tag{3.189}$$

The last one follows directly from the Lorentz representation properties of σ_{IJ}, the first one from the fact that Hermitian conjugation in combination with commuting it with γ^0 changes the sign of all σ_{IJ}.

Hermitian conjugation combined with multiplication with γ^0 is an often used operation, denoted as $\overline{\Psi} = i\Psi^\dagger\gamma^0 = i\,(\Psi^*)^T\,\gamma^0$ for spinors and $\overline{A} = -\gamma^0 A^\dagger\gamma^0$ for matrices. (Extra factors of i compared with the notation often used in particle physics appear here due to the signature.) This form of conjugation ensures that $\overline{AB} = \overline{B}\,\overline{A}$ and $\overline{A\Psi} = \overline{\Psi}\,\overline{A}$. For the matrices encountered so far, we have that $\overline{\gamma^I} = -\gamma^I$ and $\overline{\sigma_{IJ}} = -\sigma_{IJ}$.

While the matrices γ_I do not form a vector, they appear in vector currents constructed from spinor fields: the real-valued field $i\overline{\Psi}\gamma^I\Psi$ transforms as a Lorentz vector, while $\overline{\Psi}\Psi$ is scalar. To see this, note, for instance, that $i\overline{\Psi}'\gamma^I\Psi' = -\Psi^\dagger S^\dagger\gamma^0\gamma^I S\Psi = -\Psi^\dagger\gamma^0 S^{-1}\gamma^I S\Psi = i\Lambda^I{}_J\overline{\Psi}\gamma^J\Psi$. Another useful matrix constructed from the γ_I is $\gamma^5 := i\gamma^0\gamma^1\gamma^2\gamma^3$, in the representation used here given by $\gamma^5 = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}$ (seen easily using $-i\sigma_1\sigma_2\sigma_3 = \mathbb{I}$). This matrix is Hermitian, satisfies $\overline{\gamma^5} = -\gamma^5$ and anticommutes with all the initial gamma matrices. In particular, it anticommutes with γ^0, which implements a parity transformation on spinors by $\Psi' = i\gamma^0\Psi$ (so that the Dirac equation

$(\gamma^I\partial_I + m)\Psi = 0$ on Minkowski space is invariant under $x^i \mapsto -x^i$, $x^0 \mapsto x^0$, $\Psi \mapsto i\gamma^0\Psi$). Inserting an additional γ^5 in the expression for the fermion density $\overline{\Psi}\Psi$ or the fermion current $i\overline{\Psi}\gamma^I\Psi$ thus results in a pseudoscalar $\overline{\Psi}\gamma^5\Psi$ or a pseudovector $i\overline{\Psi}\gamma^5\gamma^I\Psi$, called the axial current.

On a curved space-time, we obtain tensorial fields from spinors by using a tetrad e^a_I to map internal Minkowski-space quantities to tangent-space objects. The vectorial fermion current, for instance, is $ie^a_I\overline{\Psi}\gamma^I\Psi$. The Lorentz generators σ_{IJ} are used to define a covariant SO(3,1)-derivative on spinor fields by

$$\mathcal{D}_a\Psi = \partial_a\Psi + \frac{1}{2}\omega_a^{IJ}\sigma_{IJ}\Psi = \partial_a\Psi + \frac{1}{4}\omega_a^{IJ}\gamma_{[I}\gamma_{J]}\Psi \tag{3.190}$$

with a Lorentz connection ω_a^{IJ} as introduced in Chapter 3.5.1.1.

The covariant derivative, finally, allows us to define a generally covariant fermion action, whose kinetic term is

$$S_{\text{Dirac}} = \frac{1}{2}\int_M \mathrm{d}^4x\ |e|\left(\overline{\Psi}\gamma^I e^a_I \mathcal{D}_a\Psi - \overline{\mathcal{D}_a\Psi}\gamma^I e^a_I\Psi\right)\,. \tag{3.191}$$

This action provides minimal coupling to gravity when added to the first-order Palatini action (corresponding to the Holst action for an infinite Barbero–Immirzi parameter). For the general Holst action with Barbero–Immirzi parameter γ, minimal coupling of fermions to gravity is achieved by the action

$$S^{\gamma}_{\text{Dirac}} = \frac{1}{2}\int_M \mathrm{d}^4x\ |e|\left(\overline{\Psi}\gamma^I e^a_I\left(1-\frac{i}{\gamma}\gamma^5\right)\mathcal{D}_a\Psi - \overline{\mathcal{D}_a\Psi}\left(1-\frac{i}{\gamma}\gamma^5\right)\gamma^I e^a_I\Psi\right)\,. \tag{3.192}$$

This action, as explicitly shown by Mercuri (2006), is equivalent to fermions minimally coupled to the Palatini action. (If the value of γ used here differs from that used in the Holst action for gravity, fermions are coupled non-minimally. Although pseudo-vectors appear in the action, parity is not violated at the level of equations of motion.)

3.6.4.1 *Torsion*

In contrast to the electromagnetic or scalar fields, fermionic matter couples directly to the space-time connection via the covariant derivative. This has interesting consequences in a first-order formulation of gravity coupled to fermions, since variations by the connection components, as used in part of the field equations, now receive contributions from matter terms. (A second-order formulation, by contrast, works always with the tetrad-compatible connection, and so its equations differ from the first-order ones when fermions are present.) The connection-dependent terms in the Dirac action are obtained in terms of fermion currents as

$$\frac{1}{4}e^a_I\omega_a^{JK}\left(\overline{\Psi}\gamma^I\left(1-\frac{i}{\gamma}\gamma^5\right)\sigma_{JK}\Psi + \overline{\Psi}\sigma_{JK}\left(1-\frac{i}{\gamma}\gamma^5\gamma^I\right)\Psi\right)$$

whose combinations of gamma matrices can be written as

$$\gamma^I\sigma_{JK} + \sigma_{JK}\gamma^I = \gamma^I\gamma_J\gamma_K = i\epsilon^I{}_{JKL}\gamma^5\gamma^L$$

derived using the fact that the original expression is antisymmetric in all indices, not just in J and K, and

$$-i\gamma^I\gamma^5\sigma_{JK} - i\sigma_{JK}\gamma^5\gamma^I = i\gamma^5(\gamma^I\sigma_{JK} - \sigma_{JK}\gamma^I) = 2i\gamma^5\gamma_L\delta^I_{[J}\delta^L_{K]}$$

in which the middle expression is non-vanishing only if $I = J$ or $I = K$. (This observation allows us to write $\gamma^I\sigma_{JK} - \sigma_{JK}\gamma^I = X\delta^I_{[J}\delta^L_{K]}\gamma_L$ with a number X which can be determined to equal $X = 2$ after contracting the equation with δ^J_I.) Combined with the gravitational tetrad action, which in vacuum provided (3.145), we obtain a connection variation

$$\begin{aligned}\mathcal{D}_a\left(ee^{[a}_I e^{b]}_J\right) &= 2\pi G|e|(P^{-1})_{IJ}{}^{KL}\left(e^b_I\epsilon^I{}_{KLJ}J^J + \frac{2}{\gamma}e^b_K J_L\right)\\ &= 2\pi G|e|e^b_K\epsilon^K{}_{IJ}{}^L J_L \qquad (3.193)\end{aligned}$$

(independent of γ) with the axial fermion current $J^L = i\overline{\Psi}\gamma^5\gamma^L\Psi$. In particular, the connection is no longer compatible with the tetrad. There will be a torsion contribution to the Lorentz connection, which we can write as $\omega^{IJ}_a = \widetilde{\omega}[e]^{IJ}_a + C^{IJ}_a$ with the torsion-free connection $\widetilde{\omega}^{IJ}_a$ determined by the tetrad. Equation (3.193) can then be solved to yield

$$C_{aJK} = 2\pi G e^I_a\epsilon_{IJKL}J^L\,. \qquad (3.194)$$

Canonical decomposition In a canonical analysis, the connection coupling implies that the second-class constraints, which followed from variations of the action by connection components, change in the presence of fermions. Looking first at terms containing time derivatives of the fermions, we write

$$\begin{aligned}&-\frac{i}{2}\int_{\Sigma_t}\mathrm{d}^3x\sqrt{\det h}\left((1+i/\gamma)\psi^\dagger\dot{\psi} - (1-i/\gamma)\dot{\psi}^\dagger\psi\right)\\ &= \int_{\Sigma_t}\mathrm{d}^3x\left(p_\psi\dot{\psi} - 2\pi i G(1-i/\gamma)\gamma\psi^\dagger\psi e^i_c\dot{P}^c_i\right) - \int_{\Sigma_t}\mathrm{d}^3x\frac{1-i/\gamma}{2}\mathcal{L}_t(p_\psi\psi)\,. \qquad (3.195)\end{aligned}$$

with $p_\psi = -i\sqrt{\det h}\psi^\dagger$. We will discuss the additional term containing $\dot{P}^c_i$ a little later, and for now note that the canonical structure of fermions is given by the pairs $(\psi, -i\sqrt{\det h}\psi^\dagger)$ and similarly $(\eta, -i\sqrt{\det h}\eta^\dagger)$ via 2-spinors. The axial 4-current J^I splits into its spatial components $J^i = \psi^\dagger\sigma^i\psi + \eta^\dagger\sigma^i\eta$ and the time component $J^0 = \psi^\dagger\psi - \eta^\dagger\eta$.

Performing the canonical analysis of the Holst–Dirac system along the lines of the vacuum case shows that the Ashtekar–Barbero connection $A^i_a = \Gamma^i_a + \gamma K^i_a$ receives torsion contributions to both its terms, coming from the spin connection and extrinsic curvature. (Thus, K^i_a will no longer be purely extrinsic curvature contracted with the triad.) Solving the second-class constraints as before shows that the spin-connection part changes to

$$\Gamma^k_b = \widetilde{\Gamma}^k_b - 2\pi G e^k_b J^0\,, \qquad (3.196)$$

with the torsion-free spin connection $\widetilde{\Gamma}^k_b$ obtained earlier. The torsion contribution is indeed the same one as expected from spatially projecting the space-time contribution C^{IJ}_μ to torsion:

$$C^j_a = \frac{1}{2} h^b_a \epsilon^{IJ}{}_{KL} n_I C_b{}^{KL} = -2\pi G e^j_a J^0$$

from (3.194).

While the torsion contribution to the spin connection follows directly from the second-class constraints, for the complete torsion contribution to extrinsic curvature the equations of motion for the triad must be used; the second-class constraints only tell us what additional contributions to the antisymmetric part of K_{ab} arise. The Gauss constraint now reads

$$\mathcal{D}^{(A)}_b P^b_i = \frac{1}{2}\sqrt{\det h}\, J_i \tag{3.197}$$

(including the torsion contributions in $\mathcal{D}^{(A)}_a$), and in lieu of $S_i = 0$ we have

$$\gamma \epsilon_{kmn} K^m_b P^b_n = -\sqrt{\det h}\, J_k \tag{3.198}$$

and $K^i_b e_{ai}$ can no longer be a symmetric tensor. These equations taken together imply that $\mathcal{D}_b P^b_j = 0$ even in the presence of torsion generated by the fermions, which is consistent with a spin connection (3.196) whose torsion contribution is proportional to the co-triad: $\epsilon^{ij}{}_k \Gamma^k_b P^b_j = \epsilon^{ij}{}_k \widetilde{\Gamma}^k_b P^b_j$. We write the Ashtekar–Barbero connection as

$$A^i_a = \widetilde{\Gamma}^i_a + \gamma K^i_a - 2\pi G e^i_a J^0\,. \tag{3.199}$$

3.6.4.2 *Half-densities*

Fermions also provide an interesting example for the use of fractional density weights, in this case weight $1/2$. If we couple fermions to gravity in a Holst formulation, the symplectic terms arise from (3.195). The last term is a total time derivative and can be dropped in the action, and the first one is the usual one providing us with the expression for the momentum $p_\psi = -i\sqrt{\det h}\psi^\dagger$ of ψ. But from the second term of the first integral, involving $\dot{P}^a_i$, we are led to conclude that the connection A^i_a in the combined system of gravity with fermions acquires a complex correction term $2\pi i G(1 - i/\gamma)\gamma\psi^\dagger\psi e^i_a$. One can complexify the theory and work with SL(2, $\mathbb{C}$) rather than SU(2)-connections. Already in the vacuum case, this may even be used to simplify the constraints. However, reality conditions will have to be imposed to ensure that the imaginary part of A^i_a is not arbitrary but of the required form depending on ψ and e^i_a, and those conditions are often difficult to implement.[9]

[9] The problem becomes especially pressing when one tries to quantize the theory where non-compact structure groups such as SL(2, $\mathbb{C}$) (as opposed to SU(2)) make current background-independent techniques inapplicable; see the last chapter. There is, in fact, a second problem when it comes to quantizing the fermion fields: so far, our canonical pair is (ψ, p_ψ) with $p_\psi = -i\sqrt{\det h}\psi^\dagger$. Quantization requires us to find operators on a Hilbert space such that the canonical Poisson brackets between the pair become a non-vanishing anticommutator — see the end of this section — equaling $i\hbar$, and that they commute with all other basic operators. One must also implement the reality condition $p^\dagger_\psi = i\sqrt{\det h}\psi$ in an operator form $\hat{p}^\dagger_\psi = i\widehat{\sqrt{\det h}}\hat{\psi}$. But then, for an arbitrary function $f(A)$ of the connection conjugate to the spatial metric, we have

$$0 = ([\hat{p}_\psi, f(A)])^\dagger = i[\widehat{\sqrt{\det h}}, f(A)]\hat{\psi} \neq 0 \tag{3.200}$$

which is inconsistent.

This problem, as well as difficulties with possible quantum theories, was recognized and elegantly solved by Thiemann (1998a), proposing to use half-densitized spinor fields, i.e. work with $\xi := \sqrt[4]{\det h}\psi$ instead of ψ (and $\chi := \sqrt[4]{\det h}\eta$ instead of η) as the fermionic configuration variables with momentum $p_\xi = -i\xi^\dagger$. (For unrelated reasons, half-densitized spinor fields were also introduced in the canonical formulation of supergravity by Deser et al. (1977).) The inconsistency in (3.200) is eliminated, since the reality condition $p_\xi^\dagger = -i\xi$ for half-densities does not involve the metric. Moreover, the symplectic term becomes

$$-\frac{i}{2}\int_{\Sigma_t} \mathrm{d}^3x\sqrt{\det h}\left((1+i/\gamma)(\psi^\dagger\dot{\psi} - \dot{\eta}^\dagger\eta) - (1-i/\gamma)(\dot{\psi}^\dagger\psi - \eta^\dagger\dot{\eta})\right)$$
$$= -i\int_{\Sigma_t}\mathrm{d}^3x\left(\xi^\dagger\dot{\xi} + \chi^\dagger\dot{\chi}\right) + 2\pi G\int_{\Sigma_t}\mathrm{d}^3x\, P_i^a\mathcal{L}_t(e_a^i J^0) \tag{3.201}$$

(ignoring total time derivatives). Combining the last term here with the gravitational symplectic term, there is still a contribution $2\pi G e_a^i J^0$ to be added to the Ashtekar-Barbero connection A_a^i, but it is real-valued. The new connection

$$\mathcal{A}_a^i := A_a^i + 2\pi G e_a^i J^0 = \widetilde{\Gamma}_a^i + \left(C_a^i + 2\pi G e_a^i J^0\right) + \gamma K_a^i := \widetilde{\Gamma}_a^i + \gamma K_a^i$$

can easily be employed without new reality conditions. Using half-densities has the interesting effect that the new canonical connection $\mathcal{A}_a^i$ contains only the torsion-free spin connection, even in the presence of fermions (while there is still a torsion contribution to the extrinsic-curvature part). This feature, however, is realized only for minimal coupling.

The canonical structure of gravity with Dirac fermions is thus given by the pair $(\mathcal{A}_a^i, P_j^b)$ for the gravitational variables and two pairs (ξ, p_ξ) and (χ, p_χ) of half-densitized 2-spinors. For the fermionic Poisson brackets, the canonical transformations are generated provided that one treats the spinors as Grassmann variables, i.e. their components (as well as derivatives of field-dependent functions by their components as they feature in the Poisson bracket) anticommute: $\kappa\lambda = -\lambda\kappa$ if both κ and λ are Grassmannian variables such as the components of ξ, while a fermionic component commutes with any non-fermionic field. The full motivation is due to quantum field theory, but a classical formulation of anticommuting fermions is possible. For functions f and g of the fermion components,

$$\{f, g\} = \sum_\alpha\left(\frac{\partial f}{\partial\xi_\alpha}\frac{\partial g}{\partial p_\xi^\alpha} + \frac{\partial f}{\partial p_\xi^\alpha}\frac{\partial g}{\partial\xi_\alpha}\right) \tag{3.202}$$

with a symmetric Poisson tensor (satisfying a graded Jacobi identity). Thus, $\{f, g\} = -\{g, f\}$ if f and g are Grassmann-even functions of the spinor components (functions commuting with the spinor components, for instance $\xi^\dagger\xi$). For Grassmann-odd functions κ and λ anticommuting with the spinor components (such as the spinor components themselves), $\{\kappa, \lambda\} = \{\lambda, \kappa\}$. Martin (1959) provides more details about the mathematical formulation of classical mechanics with fermionic degrees of freedom.

Example 3.26 (Gauss transformation of 2-spinors)
Using the Gauss constraint (3.197), SU(2)-gauge transformations of the half-densitized fermions are generated by the current $\frac{1}{2}\sqrt{\det h} J_j = \frac{1}{2} i(p_\xi \sigma_j \xi + p_\chi \sigma_j \chi)$. *Infinitesimal gauge transformations of the first 2-spinor are* $\{\xi, J_j\} = \frac{1}{2} i \sigma_j \xi$ *and* $\{p_\xi, J_j\} = -\frac{1}{2} i p_\xi \sigma_j$. *The minus sign in the latter expression arises from anticommuting the* p_ξ *(or the derivative by* ξ *in the Poisson bracket) past the* p_ξ *in* J_j. *The signs of these two transformations are consistent with the relationship* $p_\xi = -i\xi^\dagger$.

Exercises

3.1 Verify that the tensor $\mathcal{P}^{ij}$ defined in Example 3.3 is a Poisson tensor.

3.2 Verify the formula

$$\varepsilon^{i_1 \dots i_n} M_{i_1}^{j_1} \cdots M_{i_n}^{j_n} = \varepsilon^{j_1 \dots j_n} \det M$$

for an $n \times n$ matrix M_i^j using induction, where $\varepsilon^{i_1 \dots i_n}$ is antisymmetric in all indices with $\varepsilon^{1 \dots n} = 1$.

3.3 Compute the determinant $\det g$ of a space-time metric g_{ab} with line element

$$\mathrm{d}s^2 = -N^2 \mathrm{d}t^2 + h_{ab}(\mathrm{d}x^a + N^a \mathrm{d}t)(\mathrm{d}x^b + N^b \mathrm{d}t)$$

using the formula from the preceding exercise.

3.4 Show that the Lie derivative $\mathcal{L}_n h_{ab}$ of the spatial metric h_{ab} along the unit normal $n^a \nabla_a$ of a spatial slice is spatial, i.e. $n^a \mathcal{L}_n h_{ab} = 0$. Use only the definition of a Lie derivative but not the relation to K_{ab}.

3.5 Show that $\mathcal{L}_t T_a$ is spatial for a spatial tensor T_a if the time-evolution vector field t^a has a vanishing Lie bracket with the normal to hypersurfaces.

3.6 Let $F[h_{ab}, p^{ab}]$ be a functional of the spatial metric and its momentum. Show that the diffeomorphism constraint $C_a^{\mathrm{grav}} = -2h_{ac} D_b p^{cb}$ generates diffeomorphisms in the sense that

$$\left\{ F, \int \mathrm{d}^3x N^a(x) C_a^{\mathrm{grav}} \right\}$$
$$= \int \mathrm{d}^3x \left(\frac{\delta F}{\delta h_{ab}(x)} \mathcal{L}_N h_{ab}(x) + \frac{\delta F}{\delta p^{ab}(x)} \mathcal{L}_N p^{ab}(x) \right)$$

provided that h_{ab} and p^{ab} satisfy the constraint. Ignore boundary terms whenever you integrate by parts. It helps to consider first functionals $G[h_{ab}]$ and $H[p^{ab}]$ only of the spatial metric or its momentum, respectively, and then combine the results.

3.7 Use curvature relations to show that the vacuum constraints of canonical general relativity are proportional to the normal projections $G_{ab} n^a n^b$ and $G_{bc} h_a^b n^c$, respectively, of the Einstein tensor.

Then show that the gravitational constraints (divided by $\sqrt{\det h}$) are linear combinations of the Einstein-tensor components G_a^0 which were found to depend only on first-order time derivatives of g_{ab}. Here, an index 0 refers to projection along the time-evolution vector field.

3.8 Show that the normal-projected stess-energy components $T_{ab}n^a n^b$ and $T_{ab}n^a h^b_c$ equal the matter contributions to the constraints C and C_c (divided by $\sqrt{\det h}$).

3.9 Complete the canonical analysis of higher-derivative Lagrangians as begun in Example 3.11.

3.10 Derive the Poincaré algebra from the hypersurface-deformation algebra.

3.11 Compute the Lie derivative of g^{ab} along a vector field ξ^a, and read off the coordinate transformations of the lapse function N and the shift vector N^a under $x^\mu \mapsto x^\mu + \xi^\mu$.

Verify that the gauge transformation (3.93) generates the same transformation, observing the relationship (3.96).

3.12 Compute $\Gamma_a = g^{bc}(\partial_c g_{ab} - \frac{1}{2}\partial_a g_{bc})$ for the canonical form of the space-time metric in terms of N, N^a and h_{ab} and verify that

$$N^2 n^a \Gamma_a = N(\Gamma_0 - N^a\Gamma_a) = \partial_0 N - N^a \partial_a N - N^2 K^a{}_a$$

$$h^{ab}\Gamma_b = -D^a \log N - \frac{1}{N^2}\partial_0 N^a + \frac{1}{N^2} N^b \partial_b N^a + G^a$$

where $G^a = h^{bc} G^a{}_{bc}$ is obtained by contracting the spatial connection components.

3.13 (i) Use the canonical equations to show that the maximal-slicing condition $K^a{}_a = 0$ implies that the elliptic equation $D^a D_a N = N\,{}^{(3)}\!R$ for N.

(ii) As a simple model for a strong curvature region, assume that the spatial metric is flat while the curvature scalar is given by ${}^{(3)}\!R(r) = R_0\theta(r_0 - r)$ with constant R_0, r_0 and the step function $\theta(x)$ which is zero for $x < 0$. Solve the elliptic equation for N with these fields and show that N decreases in regions of positive scalar curvature.

Verify that the model is self-consistent for $R_0 r_0^2 \ll 1$ by solving the elliptic equation with the same scalar curvature but a spherically symmetric metric $\mathrm{d}s^2 = (1 + \delta L(r))^2 \mathrm{d}r^2 + r^2 \mathrm{d}\Omega^2$ perturbatively in δL to first order, such that ${}^{(3)}\!R(r)$ corresponds to this metric.

3.14 (i) Verify that the distortion tensor defined by

$$\Sigma_{ab} = \frac{1}{2}(\det h)^{1/3} \frac{\mathrm{d}}{\mathrm{d}t} \frac{h_{ab}}{(\det h)^{1/3}}$$

is related to shear by

$$\Sigma_{ab} = N(K_{ab} - \tfrac{1}{3}h_{ab}K^c{}_c) + D_{(a}N_{b)} - \tfrac{1}{3}h_{ab}D_c N^c\,.$$

(ii) Show that the integrated distortion $\int_\Sigma \mathrm{d}^3x \sqrt{\det h}\,\Sigma_{ab}\Sigma^{ab}$ is minimized by a shift vector satisfying $D^a\Sigma_{ab} = 0$ and express this equation as an elliptic equation

$$D_a D^a N_b + \tfrac{1}{3} D_b D_a N^a + {}^{(3)}\!R_{ab} N^a = -2D^a(N(K_{ab} - \tfrac{1}{3}h_{ab}K^c{}_c))$$

for N^a.

3.15 Use the tetrad method to compute the Ricci scalar (2.4) for an isotropic space-time with general lapse function N.

3.16 Use the tetrad method to compute the curvature invariant $R^{abcd}R_{abcd}$ for the Schwarzschild space-time.

[Hint: Eq. (5.74).]

3.17 Starting with $\{K_a^i(x), E_j^b(y)\} = 8\pi G\delta_a^b\delta^i_j\delta(x, y)$ and the relationships between K_a^i and p^{cd} and E_j^b and h_{ab}, respectively, show that the correct Poisson brackets $\{h_{ab}(x), p^{cd}(y)\} = \delta^c_{(a}\delta^d_{b)}\delta(x, y)$ for the ADM variables are produced.

3.18 Compute the constraint algebra $\{D[N_1^a], D[N_2^a]\}$, $\{H[N], D[N^a]\}$, $\{H[N_1], H[N_2]\}$ in complex Ashtekar variables.

3.19 Analyze the Holst action in Euclidean signature, replacing the Minkowski metric η_{IJ} with δ_{IJ}.

3.20 Derive the Eulerian energy current J_a^{E} and the Eulerian stress tensor S_{E}^{ab} in terms of components of T_{ab} as in Example 3.24.

3.21 Derive all Eulerian stress-energy components for the minimally coupled scalar field, and compare with the perfect-fluid form of stress-energy tensors.

3.22 Verify that the diffeomorphism constraint for half-densitized fermions generates Lie derivatives of the fermions.

3.23 Compute the canonical stress-energy components for Dirac fermions.

4

Model systems and perturbations

Many of the new phenomena predicted by general relativity are best illustrated and analyzed not by strenuous applications of the general equations but by physically motivated reductions to more special and simpler situations. One of the main tools is symmetry reduction, the investigation of solutions exhibiting certain space-time symmetry properties. The most well-known examples are spatially homogeneous geometries for cosmological models (implementing the Copernican principle) and spherically symmetric ones for non-rotating black holes. Deviations from these exactly symmetric situations are described by perturbation theory, which is heavily used, for instance, in scenarios of cosmological structure formation.

4.1 Bianchi models

The vacuum Hamiltonian constraint (3.63) can be written as

$$C_{\rm grav} = 16\pi G \mathcal{G}_{abcd} p^{ab} p^{cd} - \frac{\sqrt{\det h}}{16\pi G}\, {}^{(3)}R \tag{4.1}$$

with the *DeWitt supermetric*

$$\mathcal{G}_{abcd} = \frac{1}{\sqrt{\det h}} (h_{a(c} h_{d)b} - \tfrac{1}{2} h_{ab} h_{cd})\,. \tag{4.2}$$

It takes the form of a Hamiltonian with a kinetic term quadratic in momenta and a potential term $-\sqrt{\det h}\, {}^{(3)}R/16\pi G$. In fact, for the choice $N^a = 0$, $N = 1$ (Eulerian observers) one can easily verify that this Hamiltonian $C_{\rm grav}$ generates the equations of motion previously derived, such that in this case evolution can be interpreted as motion on the space of metrics governed by the curvature potential. (Without specifying lapse and shift, this motion is governed by $NC_{\rm grav} + N^a C_a^{\rm grav}$. Also in an inhomogeneous neighborhood of isotropic models, the Hamiltonian constraint equation takes the form of a hyperbolic evolution equation as shown by Giulini (1995).)

The space of metrics in general is an infinite-dimensional space, but is finite-dimensional when spatial homogeneity is imposed. Gravitational dynamics then provides interesting and tractable dynamical systems representing cosmological evolution.

4.1.1 Bianchi classification

Spatial homogeneity implies that there is a symmetry group S acting on space-time together with a time function t such that S acts transitively on each spatial slice Σ_t: for any pair (x, y) of points in Σ_t, there is a group element $s \in S$ such that $s(x) = y$. Moreover, on Σ_t the group action is an isometry: $s^* h_{ab} = h_{ab}$ and $s^* K_{ab} = K_{ab}$ for the action of S restricted to Σ_t such that the induced metric and extrinsic curvature, specifying the hypersurface geometry of Σ_t in space-time, are preserved.

The symmetry group must be continuous, forming a *Lie group* if its action is also required to be differentiable. The action of each element s in the connected component of the identity can be written as $s(x) = \Phi_t^\xi(x)$ with a vector field ξ^a and some $t \in \mathbb{R}$, i.e. as the action of the flow generated by ξ^a. All other values of t provide group elements as well, in particular, small ones for which we can expand $\Phi_t^\xi(x)^\mu = x^\mu + t\xi^\mu$. Since Φ_t^ξ preserves the metric, ξ^a must be a Killing vector field.

Example 4.1 (Expansion of Φ_t^ξ)
To second order in t, we write $\Phi_t^\xi(x)^\mu = x^\mu + t\xi^\mu + t^2 Y^\mu + O(t^3)$ with an object Y^μ to be determined. Since the flow is obtained by integrating along ξ^a, the identity $\Phi_t^\xi \circ \Phi_t^\xi = \Phi_{2t}^\xi$ must be satisfied: the left-hand side just splits the integration region of the right-hand side in two. (More generally, Φ_t^ξ for a fixed ξ^a is a group homomorphism from the real numbers to a 1-parameter subgroup of S.) Computing both sides with our second-order ansatz, we obtain

$$\begin{aligned}
\Phi_t^\xi(\Phi_t^\xi(x))^\mu &= \Phi_t^\xi(x^\mu + t\xi^\mu(x^\nu) + t^2 Y^\mu(x^\nu) + O(t^3)) \\
&= (x^\mu + t\xi^\mu + t^2 Y^\mu) + t\xi^\mu(x^\nu + t\xi^\nu) + t^2 Y^\mu + O(t^3) \\
&= x^\mu + 2t\xi^\mu + t^2(2Y^\mu + \xi^\nu \partial_\nu \xi^\mu) + O(t^3)
\end{aligned}$$

taking into account the position-dependence of the vector field ξ^μ, and

$$\Phi_{2t}^\xi(x)^\mu = x^\mu + 2t\xi^\mu + 4t^2 Y^\mu + O(t^3)\,.$$

Comparison results in $Y^\mu = \frac{1}{2}\xi^\nu \partial_\nu \xi^\mu$, and thus

$$\Phi_t^\xi(x)^\mu = x^\mu + t\xi^\mu + \frac{1}{2}\xi^\nu \partial_\nu \xi^\mu + O(t^3)\,. \tag{4.3}$$

Proceeding in this way for higher orders, the series expansion

$$\Phi_t^\xi(x)^\mu = \sum_{n=0}^{\infty} \frac{1}{n!} t^n (\xi^\nu \partial_\nu)^n x^\mu =: (\exp(t\xi)x)^\mu \tag{4.4}$$

results.

A Lie group can thus be interpreted as obtained by exponentiating the action of vector fields. The dimension of the space spanned by the set $\{\xi_I^a\}$ of vector fields used determines the dimension of the group. For the exponentials to form a group, certain identities must be realized between the vector fields ξ_I^a. The most important one can be seen by considering *the product commutator $\Phi_t^{\xi_1}\Phi_r^{\xi_2}\Phi_{-t}^{\xi_1}\Phi_{-r}^{\xi_2}$ for two different vector fields ξ_1^a and ξ_2^a. Inserting all expansions to second order in the parameters t and r, one obtains*

$$\Phi_t^{\xi_1}\Phi_r^{\xi_2}\Phi_{-t}^{\xi_1}\Phi_{-r}^{\xi_2}(x)^\mu = x^\mu + rt[\xi_1, \xi_2]^\mu + \cdots \tag{4.5}$$

with the vector-field commutator $[\xi_1, \xi_2]^a = \xi_1^b \partial_b \xi_2^a - \xi_2^b \partial_b \xi_1^a$. Given the group property, we know that there must be a vector field ξ_3^a and a real number q such that $\Phi_t^{\xi_1}\Phi_r^{\xi_2}\Phi_{-t}^{\xi_1}\Phi_{-r}^{\xi_2} = \Phi_q^{\xi_3}$. Comparing the expansions of both sides then shows that for any pair of vector fields ξ_1^a and ξ_2^a, there must be another vector field ξ_3^a such that $[\xi_1, \xi_2]^a \propto \xi_3^a$ and ξ_3^a gives rise to a group element when exponentiated.

A Lie group is obtained by exponentiating a set of vector fields that form a closed algebra undertaking Lie brackets. Such a set forms a *Lie algebra* defined by relations of the form

$$[\xi_I, \xi_J] = -\tilde{C}^K{}_{IJ}\xi_K \tag{4.6}$$

with *structure constants* $\tilde{C}^K{}_{IJ} \in \mathbb{R}$ antisymmetric in I and J. A Lie group that acts as the transitive symmetry group of a 3-space Σ, as considered in the present context, has a 3-dimensional Lie algebra of spatial Killing vector fields with basis $(\xi_I^a)_{I=1,\ldots,3}$. Due to transitivity, at each point x the triple $\xi_I^a(x)$ of vector fields forms a basis of the tangent space $T_x\Sigma$. As in this example, the bracket operation has arisen as the Lie bracket of vector fields; it may now be abstracted to denote the Lie bracket for composition of Lie algebra elements. By definition, it must have the properties of linearity in both entries, antisymmetry, and the Jacobi identity $\epsilon^{IJK}[\xi_I, [\xi_J, \xi_K]] = 0$. All this is automatically satisfied for a Lie bracket arising from Lie commutators of vector fields.

A Lie algebra is characterized by its structure constants $\tilde{C}^K{}_{IJ}$, and it contains much information about the Lie group obtained from it by exponentiation. Moreover, a linear algebra is easier to analyze than a non-linear group; the algebraic viewpoint is thus of advantage, focusing on properties of the structure constants. These constants cannot be chosen arbitrarily, for the relations (4.6) must obey the properties of a Lie algebra. We have already pointed out the antisymmetry of the structure constants in their two lower indices. Another relation follows from the Jacobi identity, as we will use it below for a classification of 3-dimensional Lie algebras.

Given a set of vector fields forming the Lie algebra of a transitively acting symmetry group, the Killing equation allows us to test whether a given spatial metric (or extrinsic curvature) tensor is left invariant. But it does not provide us with means to construct invariant tensors systematically, which would be required for a general analysis of homogeneous geometries and their dynamics. To that end, we will now construct invariant tensorial objects, starting with invariant vector fields which by dualization and tensorization allow us to construct invariant tensors of any type.

Invariant vector fields We assume that a Lie algebra of vector fields, together with its structure constants, has been chosen. From this, we would like to determine a tangent-space basis of invariant spatial vector fields X_I^a.

Example 4.2 (Push-forward)

So far, we have formulated the action of the Lie group on space, with coordinates x^μ. The action on vector fields or tensors is obtained by considering Φ_t^ξ as a diffeomorphism acting on the space on which the vector field is defined. For a diffeomorphism $\Phi\colon \Sigma \to \Sigma$, the usual tensor transformation law, viewing $x^\mu \mapsto \Phi^{-1}(x^\mu)^\nu$ as a change of coordinates,[1] *then provides the push-forward action*

$$\Phi_* X^\lambda(x^\mu) = \frac{\partial \Phi^{-1}(x)^\lambda}{\partial x^\nu} X^\nu(\Phi(x^\mu)) \tag{4.7}$$

of an arbitrary vector field X^a on Σ.

For a diffeomorphism Φ_t^ξ generated by a vector field ξ^a, we use the linear expansion $\Phi_t^\xi(x)^\mu = x^\mu + t\xi^\mu + O(t^2)$ in t to derive

$$\begin{aligned}\Phi_{t*}^\xi X^\lambda(x^\mu) &= (\delta_\nu^\lambda - t\partial_\nu \xi^\lambda) X^\nu(x^\mu + t\xi^\mu) + O(t^2)\\ &= X^\lambda(x^\mu) - t(X^\nu \partial_\nu \xi^\lambda - \xi^\mu \partial_\mu X^\lambda) + O(t^2)\\ &= X^\lambda(x^\mu) + t[\xi, X]^\lambda + O(t^2)\,.\end{aligned}$$

Infinitesimally, the Lie bracket determines how a diffeomorphism acts on vector fields. As discussed in more detail in the Appendix, the Lie bracket here is realized as a special case of the Lie derivative

$$\mathcal{L}_\xi X^\lambda = \lim_{t\to 0} \frac{(\Phi_{t*}^\xi - \mathbb{I}) X^\lambda}{t} = [\xi, X]^\lambda \tag{4.8}$$

applied to a vector field.

Invariant vector fields X_J^a must have vanishing Lie brackets with all symmetry generators ξ_I^a, $\mathcal{L}_{\xi_I} X_J^a = 0$, so that the symmetry flow leaves them invariant. The original vector fields ξ_I^a are themselves invariant only if $\tilde{C}^I{}_{JK} = 0$, realized only for an Abelian symmetry group. For a general symmetry type, we must construct a new set of vector fields X_I^a invariant under the flow generated by the ξ_I^a.

For a transitive group action, a basis of invariant vector fields can always be constructed, as follows by expanding $X_I^a = X_I^J \xi_J^a$ (the ξ_I^a forming a tangent-space basis) and integrating the relations

$$0 = [\xi_I, X_J]^a = (\xi_I^b \partial_b X_J^K) \xi_K^a - X_J^K \tilde{C}^L{}_{IK} \xi_L^a \tag{4.9}$$

for the functions X_I^J. According to the preceding example, this ensures that all infinitesimal actions of the Lie algebra leave the vector fields X_I^a invariant. While global integrations may be non-trivial, we can at least see the existence of local vector fields by starting at a point p where $X_I^J = (X_0)_I^J$ of some fixed choice, and then integrating the equations to a whole neighborhood in some order, e.g. first along a trajectory of ξ_1^a through p, followed

[1] If Φ transports objects such as vector fields on Σ (the active picture), their coordinate descriptions are related by Φ^{-1} (the passive picture).

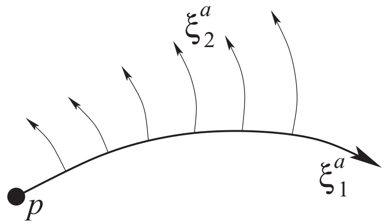

Fig. 4.1 Integration scheme for the coefficients of invariant vector fields: start at a point p, integrate along the ξ_1^a-trajectory through p, then from each point of the trajectory along ξ_2^a, and finally along ξ_3^a.

by ξ_2^a-trajectories with initial values along the ξ_1^a-trajectory, and finally for ξ_3^a (illustrated in Fig. 4.1).

Like the symmetry generators ξ_I^a, the invariant vector fields X_I^a also form a closed algebra under Lie brackets:

$$\begin{aligned}[X_I, X_J]^a &= [X_I, X_J^K \xi_K]^a = (X_I^b \partial_b X_J^K)\xi_K^a \\ &= X_I^L(\xi_L^b \partial_b X_J^K)\xi_K^a = X_I^L X_J^K \tilde{C}^M{}_{LK}\xi_M^a \\ &= C^N{}_{IJ} X_N^a \end{aligned} \tag{4.10}$$

using (4.9) and introducing

$$C^N{}_{IJ} := X_I^L X_J^K \tilde{C}^M{}_{LK}(X^{-1})_M^N\,. \tag{4.11}$$

The new structure constants $C^N{}_{IJ}$ are equivalent to the original ones because X_I^J is invertible. They are invariant under the action of the symmetry group:

$$\begin{aligned}(\xi_L^a \partial_a C^K{}_{IJ})X_K^b &= [\xi_L, C^K{}_{IJ}X_K]^b = [\xi_L, [X_I, X_J]]^b \\ &= -[X_J, [\xi_L, X_I]]^b - [X_I, [X_J, \xi_L]]^b = 0\end{aligned}$$

implies $\xi_L^a \partial_a C^K{}_{IJ} = 0$ for all L since $\{X_K^b\}$ is a basis.

In terms of Lie groups (see Appendix for more information), the two sets of commuting vector fields ξ_I^a and X_J^a generate diffeomorphisms corresponding to left- and right-translation $L_g\colon h \mapsto gh$ and $R_g\colon h \mapsto hg$ on the group manifold $S \sim \Sigma_t$. This observation demonstrates the existence of global commuting sets of vector fields because left and right multiplication commute: $R_{g_2}L_{g_1}h = (g_1 h)g_2 = g_1(hg_2) = L_{g_1}R_{g_2}h$.

Invariant tensors For a basis of invariant vector fields X_I^a as we have determined it, there is a unique dual basis of the cotangent space given by 1-forms ω_a^I satisfying $X_J^a \omega_a^I = \delta_J^I$. This relation together with the product rule and the invertibility of X_I^a already implies that the dual basis vectors are also invariant: $\mathcal{L}_{\xi_J}\omega_a^I = 0$.

As the analog of a closed Lie algebra of invariant vector fields, invariant 1-forms satisfy the *Maurer–Cartan relations*

$$D_{[a}\omega^I_{b]} = -\frac{1}{2}C^I{}_{JK}\omega^J_a\omega^K_b \tag{4.12}$$

or, in differential-form notation, $\mathrm{d}\boldsymbol{\omega}^I = -\frac{1}{2}C^I{}_{JK}\boldsymbol{\omega}^J \wedge \boldsymbol{\omega}^K$. To prove this, we note that in the X_I-basis

$$\begin{aligned}
2(D_{[a}\omega^I_{b]})X^a_J X^b_K &= X^a_J(D_a\omega^I_b)X^b_K - X^a_J(D_b\omega^I_a)X^b_K \\
&= X^a_J D_a(\omega^I_b X^b_K) - X^a_J\omega^I_b D_a X^b_K - X^a_K D_a(\omega^I_b X^b_J) + X^a_K\omega^I_b D_a X^b_J \\
&= -\omega^I_b[X_J, X_K]^b = -\omega^I_b C^L{}_{JK}X^b_L = -C^I{}_{JK}
\end{aligned}$$

using $D_a(\omega^I_b X^b_K) = 0$.

With invariant vector fields and a dual basis of invariant 1-forms, we define *invariant tensors*, of arbitrary degree, which have vanishing Lie derivatives along the ξ^a_I:

$$T^{a_1\ldots a_k}{}_{b_1\ldots b_l} = T^{I_1\ldots I_k}{}_{J_1\ldots J_l}X^{a_1}_{I_1}\cdots X^{a_k}_{I_K}\omega^{J_1}_{b_1}\cdots\omega^{J_l}_{b_l} \tag{4.13}$$

is invariant for spatial constants $T^{I_1\ldots I_k}{}_{J_1\ldots J_l}$. Invariant tensor densities of weight n are provided with an extra factor of $|\det(\omega^I_a)|^n$. In particular, we define a class of homogeneous spatial metrics

$$h_{ab} = h_{IJ}\omega^I_a\omega^J_b \tag{4.14}$$

with a symmetric matrix h_{IJ} whose coefficients depend on time only. Any such metric h_{ab} is invariant under the given group action of S, and is thus homogeneous with Killing vector fields ξ^a_I, $I = 1, 2, 3$. As can be seen for instance in Hawking (1969), the structure constants determine the curvature of any such metric by the Ricci scalar

$$^{(3)}R = -\frac{1}{2}C^I{}_{JK}C^K{}_{LI}h^{JL} - \frac{1}{4}C^I{}_{JK}C^L{}_{MN}h^{JM}h_{IL}h^{KN} - C^I{}_{IJ}C^K{}_{KL}h^{JL} \tag{4.15}$$

as it enters the dynamics via the curvature potential in the Hamiltonian constraint.

Example 4.3 (Homogeneous metric)
For an Abelian Lie algebra $\mathbb{R}^3$, the structure constants vanish and the Maurer–Cartan relations are easily solved by differentials $\omega^I_a = (\mathrm{d}x^I)_a$ of the Cartesian coordinates x^I in $\mathbb{R}^3 = \Sigma$. A dual basis of invariant vector fields is given by coordinate derivatives $X^a_I = (\partial/\partial x^I)^a$. Any invariant spatial metric obeying an $\mathbb{R}^3$-symmetry is of the form $h_{ab} = h_{IJ}(\mathrm{d}x^I)_a(\mathrm{d}x^J)_b$ with spatially constant (but possibly time-dependent) coefficients h_{IJ}. If some structure constants are non-vanishing, no invariant coordinate basis of the tangent space exists.

Structure constants So far, we have assumed a given symmetry group with its Lie algebra and determined its invariant tensors. Via invariant vector fields and 1-forms used to construct them, the form of invariant tensors depends on the structure constants. The structure constants themselves cannot be chosen arbitrarily because the axioms satisfied by the symmetry group imply relations between them.

More precisely, the symmetry generators ξ_I^a and the invariant vector fields X_I^a form a Lie algebra, which must obey antisymmetry of the Lie bracket and the Jacobi identity. Moreover, we do not want to distinguish between different versions of $C^I{}_{JK}$ just obtained by redefining the basis of symmetry generators by linear transformations. This reduces the possible choices for $C^I{}_{JK}$ to nine types, as they have been classified by Bianchi.

It is useful to parameterize the options in the following form: first, using antisymmetry, $C^I{}_{JK} = C^I_{[JK]}$, all the information in the structure constants can equivalently be expressed by a matrix

$$\frac{1}{2} C^I{}_{JK} \epsilon^{JKL} =: n^{(IL)} + A^{[IL]} = n^{IL} + \epsilon^{ILK} a_K \,.$$

In the first step, we have decomposed this matrix into its symmetric and antisymmetric parts, and then expressed the antisymmetric part A^{IL} by its three non-trivial components gathered in the vector a_K. With this decomposition we express the structure constants as

$$C^I{}_{JK} = \frac{1}{2}(C^I{}_{JK} - C^I{}_{KJ}) = \frac{1}{2} C^I{}_{LM} \epsilon^{LMN} \epsilon_{NJK} = \epsilon_{NJK} n^{IN} + \delta^I_K a_J - \delta^I_J a_K \,. \quad (4.16)$$

Finally, the symmetric matrix can be diagonalized by a constant change of the basis X_I^a as $n^{IJ} = n^{(I)} \delta^{IJ}$ (where we are not summing over I on the right-hand side, as indicated by the brackets). We are left with six independent parameters n^I and a_J in the structure constants.

The vector a_I plays an important role in the classification of Bianchi models by their structure constants because it allows one to split them in two classes, the so-called Bianchi class A models which have $a_I = \frac{1}{2} C^J{}_{IJ} = 0$, and Bianchi class B models for which $a_I \neq 0$. With the Jacobi identity

$$[X_I, [X_J, X_K]] + \text{cyclic} = 0 \,,$$

which we have not used so far in the classification, an identity for the structure constants follows, implying $n^{IJ} a_J = 0$ (Exercise 4.1). Thus, the vector a_I is an eigenvector of the symmetric matrix n^{IJ} and, without loss of generality, can be chosen to be $(a, 0, 0)$ in the eigenbasis that makes n^{IJ} diagonal. Now, only four independent parameters are left in the classification.

Redefinitions of the Lie algebra basis allow further simplifications by rescaling and reflecting basis vectors. In this way, all components can be made either zero or ± 1, and only relative signs between the parameters are relevant. The matrix n^{IJ} and the vector a_I can be interpreted as tensors on $\mathbb{R}^3$ transforming under linear transformations. Without raising or lowering indices and noting that $n^{IJ} a_J = 0$ vanishes automatically, this allows

Table 4.1 *Classification of the Bianchi types. (Type III is identical to type* $\mathrm{VI}_{\eta=-1}$*.)*

	Group type	a	n^1	n^2	n^3
Class A	I	0	0	0	0
	II	0	$+$	0	0
	VII_0	0	$+$	$+$	0
	VI_0	0	$+$	$-$	0
	IX	0	$+$	$+$	$+$
	VIII	0	$+$	$+$	$-$
Class B	V	$+$	0	0	0
	IV	$+$	0	0	$+$
	VII_η	$+$	0	$+$	$+$
	VI_η	$+$	0	$+$	$-$

only one possible linear invariant, η, defined by

$$a_I a_J = \frac{1}{2}\eta\, \epsilon_{IKL}\, \epsilon_{JMN}\, n^{KM} n^{LN}\,.$$

In an eigenbasis of n^{IJ}, we have $a^2 = \eta n^2 n^3$. Thus, η can be identified with zero in class A models, and is well defined in class B models only for $n^2 n^3 \neq 0$. Since η is invariant under linear transformations, it is the only continuous parameter, provided that it is well-defined and not restricted to vanish. Altogether, this provides a complete classification of Bianchi models corresponding to inequivalent 3-dimensional Lie algebras as summarized in Table 4.1, where subscripts on the traditional group symbols indicate the continuous freedom of η. For class B models, not all n^I can be non-zero because the matrix n^{IJ} must have a non-trivial zero eigenvector: $n^{IJ} a_J = 0$ and $a \neq 0$ in this case.

4.1.2 Diagonalization

We now turn to the dynamics of Bianchi models in general relativity. Since the coefficients h_{IJ} of a homogeneous metric (4.14) form a symmetric matrix, they can always be diagonalized by rotating the basis of invariant 1-forms ω_a^I: an orthogonal matrix $O_I{}^J$, $O_I{}^J (O^T)_J{}^K = \delta_I^K$, exists such that $O_I{}^J h_{JK} O_L{}^K = h_{(I)} \delta_{IL}$. We then write the spatial metric at any given time as

$$h_{ab} = h_{IJ}\omega_a^I \omega_b^J = (O_I{}^J h_{JK} O_L{}^K)((O^T)_M{}^I \omega_a^M)((O^T)_N{}^L \omega_b^N) = h_{(I)} \delta_{IJ} \omega_a'^I \omega_b'^J$$

with $\omega_a'^I := (O^T)_J{}^I \omega_a^J$.

Initial conditions can thus always be made diagonal by a choice of basis. Under evolution, the spatial metric stays diagonal if the momentum p^{IJ} is diagonal in the same basis at all times. As initial momenta, we can choose a diagonal form just as we did for h_{IJ}, but this will be preserved in time only if $\dot{p}^{IJ}$, or the spatial Ricci tensor related to it by the evolution equations (3.80), is diagonal. In general, this is only the case for class A models. In what follows we will restrict attention to diagonalizable models which already exhibit the most interesting features of homogeneous models, and thus assume that $a_I = 0$ together with a diagonal metric $h_{IJ}(t) = h_{(I)}(t)\delta_{IJ}$ valid at all times.

The dynamics of diagonal models is most easily formulated for a vanishing shift vector $N^a = 0$ and a lapse "function"

$$N = \frac{\sqrt{\det h}}{16\pi G} \tag{4.17}$$

(which for this choice is actually a scalar density). Since we assume class A, we have structure constants $C^I{}_{JK} = \epsilon_{IJK} n^{(I)}$. For a diagonal spatial metric $h_{IJ} = h_{(I)}\delta_{IJ}$ and vanishing shift vector, one easily computes the extrinsic curvature $K_{IJ} = \frac{1}{2N}\dot{h}_{(I)}\delta_{IJ}$ from (3.51), and thus also the momentum $p^{IJ} = p^{(I)}\delta^{IJ}$ (derived from a density-weighted homogeneous $p^{ab} = |\det(\omega_a^K)| p^{IJ} X_I^a X_J^b$) is diagonal. For its components and those of h_{IJ}, we have the Poisson bracket $\{h_I, p^J\} = \delta_I^J$, as it follows from the pull-back of the homogeneous symplectic term to the subspace of diagonal configurations:

$$\int \mathrm{d}^3x \dot{h}_{ab} p^{ab} = \int \mathrm{d}^3x |\det(\omega_a^K)| \dot{h}_{IJ} p^{IJ} = \int \mathrm{d}^3x |\det(\omega_a^K)| \dot{h}_I p^I = \dot{h}_I p^I\,.$$

(As in Chapter 2, we set the coordinate volume $V_0 = \int \mathrm{d}^3x |\det(\omega_a^K)| = 1$.)

These reductions can now be inserted into the Hamiltonian constraint

$$N C_{\mathrm{grav}} = p_{ab} p^{ab} - \frac{1}{2}(p^a{}_a)^2 - \frac{\det h}{(16\pi G)^2}\, {}^{(3)}R = 0 \tag{4.18}$$

whose curvature potential is given by (4.15). The kinetic term quadratic in momenta is

$$\begin{aligned} p_{ab}p^{ab} - \frac{1}{2}\left(p^a{}_a\right)^2 &= h_{ac}h_{bd}p^{cd}p^{ab} - \frac{1}{2}(h_{ab}p^{ab})^2 \\ &= h_{IK}h_{JL}p^{KL}p^{IJ} - \frac{1}{2}\left(h_{IJ}p^{IJ}\right)^2 \\ &= \sum_{I=1}^{3} h_I^2\left(p^I\right)^2 - \frac{1}{2}\left(\sum_{I=1}^{3} h_I p^I\right)^2 \\ &= \frac{1}{2}\left(h_1^2(p^1)^2 + h_2^2(p^2)^2 + h_3^2(p^3)^2\right) - h_1h_2p^1p^2 \\ &\quad - h_1h_3p^1p^3 - h_2h_3p^2p^3\,. \end{aligned}$$

Misner variables Even with diagonal metrics, the kinetic term is not of diagonal form in the sense that there are cross products of different momentum components. For convenience, one can change this by a canonical transformation, following Misner (1969), that we split into two steps:

1. In the first step, we do not yet diagonalize but absorb the metric components of the kinetic term into new momenta

$$\alpha_I := \frac{1}{2}\log h_I\,, \qquad \rho^I := 2h_{(I)}p^I \quad \text{such that} \quad \{\alpha_I, \rho^J\} = \delta_I^J\,. \tag{4.19}$$

The kinetic term then depends only on ρ^I.

2. Now we define the *Misner variables* $(\alpha, \beta_+, \beta_-)$ and their momenta (p_α, p_+, p_-) via

$$\alpha_1 =: \alpha + \beta_+ + \sqrt{3}\beta_-\,, \qquad \rho^1 =: \frac{1}{3}p_\alpha + \frac{1}{6}p_+ + \frac{1}{2\sqrt{3}}p_- \tag{4.20}$$

$$\alpha_2 =: \alpha + \beta_+ - \sqrt{3}\beta_-\,, \qquad \rho^2 =: \frac{1}{3}p_\alpha + \frac{1}{6}p_+ - \frac{1}{2\sqrt{3}}p_- \tag{4.21}$$

$$\alpha_3 =: \alpha - 2\beta_+\,, \qquad \rho^3 =: \frac{1}{3}p_\alpha - \frac{1}{3}p_+\,. \tag{4.22}$$

It can easily be verified that this transformation makes the kinetic term diagonal in momenta:

$$\begin{aligned} p_{ab}p^{ab} - \frac{1}{2}(p_a^a)^2 &= \frac{1}{8}\left((\rho^1)^2 + (\rho^2)^2 + (\rho^3)^2\right) - \frac{1}{4}(\rho^1\rho^2 + \rho^1\rho^3 + \rho^2\rho^3) \\ &= \frac{1}{24}(-p_\alpha^2 + p_+^2 + p_-^2)\,. \end{aligned} \tag{4.23}$$

In Misner variables, the geometry is not described by the usual metric components but by the total spatial volume

$$\alpha = \frac{1}{3}(\alpha_1 + \alpha_2 + \alpha_3) = \log(\det h_{IJ})^{1/6} \tag{4.24}$$

and the *anisotropy parameters* $\beta_\pm$ which vanish for isotropic metrics where $\alpha_1 = \alpha_2 = \alpha_3$.

Anisotropy potential The kinetic term is to be combined with the curvature potential (4.15) which we write as

$$^{(3)}R = -\frac{1}{2}\left(\frac{n^1h_1}{h_2h_3} + \frac{n^2h_2}{h_1h_3} + \frac{n^3h_3}{h_1h_2} - 2\frac{n^1n^2}{h_3} - 2\frac{n^1n^3}{h_2} - 2\frac{n^2n^3}{h_1}\right) \tag{4.25}$$

$$\begin{aligned} &= -\frac{1}{2}e^{-2\alpha}\left(n^1e^{4(\beta_+ + \sqrt{3}\beta_-)} + n^2e^{4(\beta_+ - \sqrt{3}\beta_-)} + n^3e^{-8\beta_+}\right. \\ &\quad \left. - 2n^1n^2e^{4\beta_+} - 2n^1n^3e^{-2(\beta_+ - \sqrt{3}\beta_-)} - 2n^2n^3e^{-2(\beta_+ + \sqrt{3}\beta_-)}\right) \end{aligned} \tag{4.26}$$

for diagonal metrics and then in Misner variables. This gives a curvature potential $-e^{6\alpha\ (3)}R/(16\pi G)^2$ in (4.18).

Equations of motion for the metric components and with our choice (4.17) of the lapse function are simply

$$\dot{\alpha} = \{\alpha, NC_{\mathrm{grav}}\} = -\frac{1}{12}p_\alpha\,, \qquad \dot{\beta}_+ = \frac{1}{12}p_+\,, \qquad \dot{\beta}_- = \frac{1}{12}p_- \tag{4.27}$$

which allows us to write the Hamiltonian constraint as

$$-6\dot{\alpha}^2 + 6\dot{\beta}_+^2 + 6\dot{\beta}_-^2 - \frac{e^{6\alpha\ (3)}R}{(16\pi G)^2} = 0\,.$$

4.1.3 Kasner solutions and BKL scenario

Anisotropic geometries, even if they are homogeneous, allow more general kinds of curvature than the isotropic FLRW models. They are thus especially important for an understanding of the behavior of universe models near a singularity, when curvature is large. In fact, specific solutions as well as general properties of anisotropic models show that the isotropic approach to a singularity is very special. The anisotropic approach, on the other hand, seems generic, as indicated by several arguments to be discussed now.

Bianchi I and Kasner solutions The simplest Bianchi model is obtained for type I, in which case the symmetry group is Abelian: $\mathbb{R}^3$ or a compactification. All structure constants vanish, $C^I{}_{JK} = 0$, and invariant 1-forms can simply be chosen as differentials of Cartesian coordinates: $\omega^I_a \mathrm{d}x^a = \mathrm{d}x^I$; see also Example 4.3.

As a further consequence, the curvature potential vanishes identically. The system with Hamiltonian constraint

$$NC_{\mathrm{grav}} = \frac{1}{24}(-p_\alpha^2 + p_+^2 + p_-^2) \tag{4.28}$$

becomes deparameterizable in internal time α, providing solutions $\beta_\pm(\alpha)$ with preserved momenta. With a constant potential in the anisotropy plane, solutions for $\beta_\pm(\alpha)$ are straight lines. In fact, we can easily solve equations of motion also in coordinate time: for the variables (4.19),

$$\dot{\rho}^I = \{\rho^I, NC_{\mathrm{grav}}\} = 0$$

implies that all ρ^I are constant in time, and thus also

$$\dot{\alpha}_I = \{\alpha_I, NC_{\mathrm{grav}}\} = \frac{1}{4}(2\rho^I - \rho^1 - \rho^2 - \rho^3) =: \tilde{v}^I$$

are constant. Solutions are $\alpha_I(t) = \tilde{v}^I t + \alpha_I^{(0)}$ in a time coordinate t corresponding to our choice of lapse. The constants are not arbitrary because initial values must satisfy the Hamiltonian constraint

$$0 = NC_{\mathrm{grav}} = \frac{1}{8}\sum_I (\rho^I)^2 - \frac{1}{4}(\rho^1\rho^2 + \rho^1\rho^3 + \rho^2\rho^3) = \sum_I \dot{\alpha}_I^2 - \left(\sum_I \dot{\alpha}_I\right)^2$$

which requires

$$\sum_I (\tilde{v}_I)^2 = \left(\sum_I \tilde{v}_I\right)^2 . \tag{4.29}$$

For any such solution, we transform back to the original metric components

$$h_I = e^{2\alpha_I} = h_I^{(0)} e^{2\tilde{v}^I t} ,$$

compute the lapse function $N \propto \sqrt{\det h} \propto \exp((\tilde{v}^1 + \tilde{v}^2 + \tilde{v}^3)t)$, and obtain proper time $\tau = \int^\tau N(t)\mathrm{d}t \propto \exp((\tilde{v}^1 + \tilde{v}^2 + \tilde{v}^3)t)$. In this time coordinate, the line element takes the form

$$ds^2 = -d\tau^2 + \tau^{2v^1}(dx^1)^2 + \tau^{2v^2}(dx^2)^2 + \tau^{2v^3}(dx^3)^2 \tag{4.30}$$

of the *Kasner metric* , where we absorbed the constants $h_I^{(0)}$ in the spatial coordinates and defined the *Kasner exponents*

$$v^I = \frac{\tilde{v}^I}{\sum_J \tilde{v}^J}\,.$$

They must satisfy

$$\sum_I v^I = 1 = \sum_I (v^I)^2\,, \tag{4.31}$$

the first part by definition, the second due to the Hamiltonian constraint (4.29).

Kasner solutions have several characteristic properties. The volume as a function of proper time is proportional to $\tau^{v^1+v^2+v^3} = \tau$, which vanishes at $\tau = 0$ independently of the Kasner exponents. The conditions (4.31) allow only solutions for which $-1 < v^I \leq 1$ in such a way that one of the exponents is negative while the other two are positive. Thus, it is not possible that all directions are expanding or contracting at the same time, which shows the difference to isotropic solutions. Moreover, extrinsic curvature, whose components are given by time derivatives of h_I and thus expressions proportional to τ^{2v^I-1}, does not remain finite at $\tau = 0$, while the spatial Ricci scalar vanishes. By the Gauss equation, the space-time Ricci scalar diverges at $\tau = 0$, presenting a singularity at a finite proper time: relaxing the assumption of isotropy does not eliminate singularities. (See also Exercise 4.3.)

Bianchi IX and mixmaster behavior For the Bianchi IX model, we have structure constants $C^I{}_{JK} = \epsilon^I{}_{JK}$ of the Lie algebra su(2), and a non-trivial curvature potential

$$\begin{aligned} -\frac{e^{6\alpha}}{(16\pi G)^2}\,{}^{(3)}R = \frac{1}{2}\frac{e^{4\alpha}}{(16\pi G)^2}\Big(& e^{-8\beta_+} - 4e^{-2\beta_+}\cosh(2\sqrt{3}\beta_-) \\ &+ 2e^{4\beta_+}(\cosh(4\sqrt{3}\beta_-) - 1)\Big) \end{aligned} \tag{4.32}$$

with the following properties:

1. From the deparameterized perspective used initially for Kasner solutions, $e^{6\alpha}\,{}^{(3)}R$ presents a time (α) dependent potential in the anisotropy plane (β_+, β_-). The system is no longer deparameterizable by α.
2. The potential is symmetric under rotations of (β_+, β_-) by $2\pi/3$.
3. Along radial directions away from the isotropy point $\beta_+ = 0 = \beta_-$, the potential generically has exponentially increasing walls, except for the direction $\beta_- = 0$, $\beta_+ \to \infty$ and its two rotations by $\pm 2\pi/3$.
4. The shape of the typical wall can be determined from $\beta_- = 0$, $\beta_+ \to -\infty$: its dominant increase is by $e^{8|\beta_+|}$.

Some of these features can be seen in Fig. 4.2. (See also the cover image.)

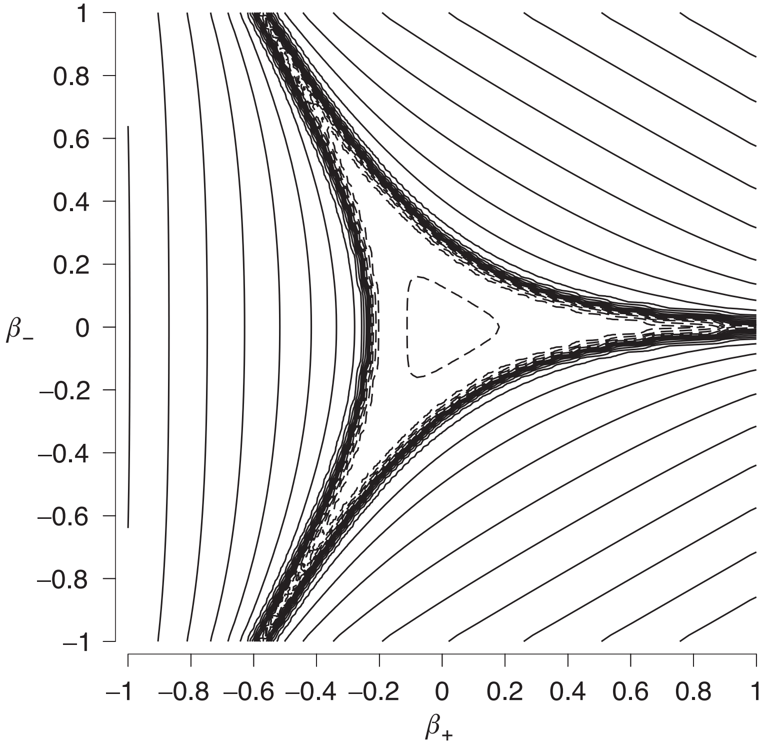

Fig. 4.2 Logarithm of the norm of the curvature potential for a Bianchi IX model in the anisotropy plane at fixed volume. Solid lines of this contour plot correspond to positive values of the potential, dashed lines to negative values.

Even though the system is not globally deparameterizable by α, we can visualize and describe the evolution by the picture of a point particle moving in the anisotropy potential. When the potential is small near the isotropy point, we have almost free Kasner motion as in the Bianchi I model, a so-called Kasner epoch. However, the particle, following a straight line, moves away from isotropy and generically approaches one of the exponential walls. Because of their steepness, reflections at the walls can be approximated as almost instantaneous. The Kasner exponents v^I change during such a reflection in a way which can be determined from the equations of motion. After a reflection, we enter a different Kasner epoch in such a way that the previously negative exponent becomes positive and one of the positive ones becomes negative. The directions change roles at each reflection regarding their expansion/contraction behavior. There are infinitely many reflections before time $t = 0$ is reached, earning the dynamics the moniker *mixmaster behavior*. At any time, the volume is decreasing in the evolution toward smaller t, and hits a singularity after a finite amount of proper time. Because reflections happen in the triangular shape of the potential which encloses a finite volume in the anisotropy plane, general results of billiard systems imply that the dynamics is chaotic; see, for instance, the review by Damour et al. (2003).

BKL scenario As first undertaken by Belinskii et al. (1982), one may look for generic asymptotic line elements close to a spacelike singularity at $\tau = 0$ by making an ansatz of the form

$$ds^2 = -d\tau^2 + \tau^{2v^1(x)}(\boldsymbol{\omega}^1)^2 + \tau^{2v^2(x)}(\boldsymbol{\omega}^2)^2 + \tau^{2v^3(x)}(\boldsymbol{\omega}^3)^2 \tag{4.33}$$

with homogeneous 1-forms $\boldsymbol{\omega}^I$ of some Bianchi type and "Kasner exponents" $v^I(x)$ which are now space-dependent. The equations of motion for a homogeneous model corresponding to the invariant 1-forms $\boldsymbol{\omega}^I$ are then modified by terms containing spatial derivatives $\partial_a v^I(x)$.

The spatial curvature tensor then acquires two contributions, one from the invariant 1-forms which may have non-vanishing derivatives, and one from the spatially dependent $v^I(x)$. The first contribution simply agrees with the curvature tensor of a homogeneous model as used before. The second contribution captures properties of the inhomogeneity and is thus more interesting in the present context. By the product rule, the spatial connection coefficients

$$G^c_{ab} = \frac{1}{2} h^{cd}(\partial_a h_{bd} + \partial_b h_{ad} - \partial_d h_{ab})$$

with $h_{ab} = \delta_{ab}\tau^{2v^{(b)}}$ from (4.33) acquire a contribution

$$\delta G^c_{ab} = \log\tau(\delta^c_b \partial_a v^{(b)} + \delta^c_a \partial_b v^{(a)} - \delta^{cd}\delta_{ab}\partial_d v^{(a)})$$

from the inhomogeneity, which diverges as $\log\tau$ at the singularity. Spatial Riemann or Ricci curvature tensors, quadratic in the connection coefficients, diverge as $(\log\tau)^2$, and the Ricci scalar has diverging contributions of the order $\tau^{-2v^I}(\log\tau)^2$ with an additional factor from the inverse metric.

There are also curvature terms not containing derivatives of v^I, which are the same as in the homogeneous case. The strongest divergence in the Ricci scalar (4.25) comes from terms of the form

$$\frac{h_1}{h_2 h_3} \propto \tau^{2v^1 - 2v^2 - 2v^3} .$$

This particular term is dominant if $v^1 < 0$ is the negative one of the exponents. Because of $2v^1 - 2v^2 - 2v^3 < -2v^I$ for all v^I if $v^1 < 0$, the homogeneous contribution to spatial curvature always diverges more strongly than the inhomogeneous contributions from spatial derivative terms of v^I. Thus, spatial derivatives are subdominant for asymptotic solutions near a space-like singularity if the metric takes the form (4.33). The local behavior is then that of a homogeneous model, and in general resembles the Mixmaster behavior of the Bianchi IX (or VIII) model since the strongest divergences arise when all n^I are non-vanishing.

BKL's conclusion is encouraging, for it suggests that the approach to generic spacelike singularities can be understood solely in terms of homogeneous models. There are, however, several caveats and the BKL conjecture, stating that the homogeneous behavior characterizes even inhomogeneous singularities, is not completely proven. (Growing numerical evidence exists, for instance by Garfinkle (2004) and reviewed by Berger (2002), as well as partial analytical support by Uggla et al. (2003), Rendall (2005) and Andersson and Rendall

(2001). Black-hole singularities have been analyzed in this spirit as well, by the numerical studies of Saotome et al. (2010).) First, one should notice that the conjecture does not at all state that space-like singularities must be homogeneous. It is just the local behavior at each point that agrees with the dynamics of a Bianchi model; the exponents v^I remain spatially dependent even though their spatial derivatives do not strongly affect the dynamics.

Moreover, the chaos of the mixmaster behavior is relevant, since it implies further fragmentation and in fact a very inhomogeneous general appearance of singularities: far from the singularity, one may start with an approximation of spatial geometries by a collection of homogeneous patches, and then evolve backwards. Nearby patches which initially were similar in their metrics differ rapidly from each other as time goes on because the chaotic dynamics makes evolution highly sensitive to initial data. To maintain the approximation by homogeneous patches, one must subdivide them. Spatial slices become more and more structured and inhomogeneous. Spatial gradients then seem to be growing, which might challenge the BKL picture.

Example 4.4 (Gowdy models)

Gowdy (1974) introduced a class of inhomogeneous models which by now serve several interesting purposes in the analysis of cosmological situations, for instance in the context of the BKL behavior. In one type of model, the metric takes the form

$$\mathrm{d}s^2 = e^{(\lambda+t)/2}(-e^{-2t}\mathrm{d}t^2 + \mathrm{d}x^2) + e^{-t}\left(e^P(\mathrm{d}y + Q\mathrm{d}z)^2 + e^{-P}\mathrm{d}z^2\right) \tag{4.34}$$

with three functions $\lambda(t,x)$, $P(t,x)$, $Q(t,x)$ of t and x only. A singularity is approached for $t \to \infty$, which amounts to finite proper time. When (4.34) is inserted into Einstein's equation, it produces the equations

$$\begin{aligned} \partial_t^2 P - e^{2P}(\partial_t Q)^2 - e^{-2t}\partial_x^2 P + e^{2(P-t)}(\partial_x Q)^2 &= 0 \\ \partial_t^2 Q + 2\partial_t P \partial_t Q - e^{-2t}(\partial_x^2 Q + 2\partial_x P \partial_x Q) &= 0 \end{aligned} \tag{4.35}$$

for its coefficients. Once solutions for P and Q are known, λ can be determined from

$$\begin{aligned} \partial_t\lambda &= -\left((\partial_t P)^2 + e^{2P}(\partial_t Q)^2\right) - e^{-2t}\left((\partial_x P)^2 + e^{2P}(\partial_x Q)^2\right) \\ \partial_x\lambda &= -2(\partial_t P\partial_x P + e^{2P}\partial_t Q\partial_t P)\,. \end{aligned}$$

As the singularity is approached, spatial derivatives are suppressed by factors $\exp(-2t)$ *while time derivatives remain.*

The BKL scenario is thus supported by Gowdy models, provided that P stays bounded as the singularity is approached for $t \to \infty$. If P diverges as well, not all the spatial derivatives are necessarily suppressed, since $e^{2(P-t)}$ appears in one coefficient in (4.35). Indeed, so-called spikes, isolated points where P grows large and gradients become steep, have been found in numerical as well as analytical solutions. Spikes seem to remain isolated and do not challenge the general BKL ideas; they do, however, illustrate some of the difficulties in finding a rigorous proof of the conjecture.

Analytical results related to the BKL picture have been reviewed by Rendall (2005), and numerical progress is described by Berger (2002), Garfinkle (2004) and Lim et al. (2009).

Spikes have been constructed explicitly and compared with numerical results by Rendall and Weaver (2001).

In the fragmentation argument, there was no obvious violation of time reversal symmetry, and thus one may even expect fragmentation to the future. Such a behavior would be in conflict with observations of structures in the universe. The contradiction is resolved by noting that fragmentation relied on the chaotic behavior of the Bianchi IX dynamics, which may be used reliably near a spacelike singularity. Evolving to the future, away from the singularity, leads to a different form of dynamics to which the asymptotic solution discussed here does not apply. To understand homogenization and isotropization in an expanding universe, one must use different approximations valid at later times.

For homogenization, inflation is often used as an argument, stating that a small patch of an early geometry is enlarged rapidly to all that we can see now. If the early patch, by virtue of being tiny, was very nearly homogeneous, our currently visible universe should still be nearly homogeneous. However, this argument is incomplete: given the preceding discussion, we expect chaotic behavior to play a large role in the dynamics of the very early universe. With chaos, one often sees fractal structures which should arise in the spatial geometries, for instance by the fragmentation process sketched above. But if a self-similar fractal is enlarged by some factor of whatever size, its structure will not change. Thus, as pointed out by Penrose (1990), inflation by itself cannot explain the homogeneity of the current universe (and may not even get started if the initial geometry is too inhomogeneous). Isotropization, once a nearly homogeneous space-time is assumed, can be explained more easily based on the behavior of Bianchi models containing perfect-fluid matter sources as first studied by Misner (1968). A simple heuristic argument uses the fact that anisotropies contribute a shear term to the Friedmann equation of an average isotropic universe, which as a contribution to energy density depends on $\dot{\beta}_\pm$ and behaves as a^{-6}: a much faster drop-off behavior than that of standard matter. Anisotropies thus do not play a dominant role for the dynamics of a large universe.

Example 4.5 (Anisotropic shear)
A universe of Bianchi type I containing matter obeys the Hamiltonian constraint

$$NC = \frac{1}{24}(-p_\alpha^2 + p_+^2 + p_-^2) + \frac{e^{6\alpha}}{16\pi G}\rho = 0$$

with the anisotropic variables used before for the vacuum model, and the energy density ρ of matter ($-16\pi G\rho$ taking the place of the spatial Ricci scalar in the Hamiltonian constraint of general Bianchi models). The gravitational part of this constraint was obtained in (4.28) with a lapse function $N = \sqrt{\det h}/16\pi G = e^{3\alpha}/16\pi G$. The derivative of α by proper time is thus $\mathrm{d}\alpha/\mathrm{d}\tau = \{\alpha, C\} = -\frac{4}{3}\pi G e^{-3\alpha} p_\alpha$. Combining this equation with the constraint and recalling that $\alpha = \log a$, we can write the dynamics in Friedmann form:

$$\left(\frac{\dot{a}}{a}\right)^2 = \left(\frac{\mathrm{d}\alpha}{\mathrm{d}\tau}\right)^2 = \frac{16\pi^2 G^2}{9}\frac{p_\alpha^2}{e^{6\alpha}} = \frac{8\pi G}{3}\rho + \frac{16}{9}\pi^2 G^2 \frac{p_+^2 + p_-^2}{a^6}\,. \tag{4.36}$$

During a Kasner epoch, the momenta p_+ and p_- are nearly constant provided that, in the presence of matter, ρ is isotropic and does not depend strongly on the anisotropy parameters $\beta_\pm$. The shear term provided by an anisotropic geometry then behaves as a^{-6}, and in an expanding universe drops off faster than the usual matter ingredients. (Anisotropies show the dilution behavior of a stiff fluid.)

4.2 Symmetry

As demonstrated by Bianchi models, Einstein's equation, also in canonical form, can often be analyzed rather easily in models exhibiting continuous space-time symmetries. Various important phenomena are illustrated in this way, and in many cases have been uncovered only thanks to sufficiently general symmetric models. A framework for the formulation and derivation of symmetry-reduced dynamical equations thus constitutes an important tool to analyze canonical gravity.

4.2.1 Symmetry reduction

We will first describe the basic constructions of symmetric models in terms of their equations of motion. In the following parts of this subsection, general classification schemes and additional ingredients as well as subtleties for the reduction of canonical models will be discussed. Throughout this section, we will use the example of symmetry reduction to develop techniques of principal fiber bundles and connections. This can only be a brief introduction; for details, we suggest reading Göckeler and Schücker (1989).

4.2.1.1 Symmetric space-times

For symmetric models, we are looking for space-time metrics g_{ab} invariant under the transformations of a certain symmetry group $S\colon M \to M$, i.e. $s^* g_{ab} = g_{ab}$ for all $s \in S$. The orbits of S in M are generated by Killing vector fields of invariant metrics.

The symmetry group always acts on space-time M. To be left with a non-stationary and thus dynamical reduced theory, however, we require that all Killing vector fields are spacelike. Derived as infinitesimal generators of a group action, they form a closed algebra under the Lie bracket, they are in involution and thus span the tangent spaces to spatial submanifolds, the spatial orbits of the symmetry groups along which invariant tensors do not change. If the orbits are 3-dimensional, homogeneity is realized with a transitive action of the symmetry group on a distinguished class of spatial 3-manifolds given by its orbits. These models, called mini-superspaces, are those of the Bianchi classification, or (with a further generator of an isotropy) the Kantowski–Sachs model.[2] If the orbits are less than 3-dimensional, we obtain a midi-superspace model. Their orbit spaces obtained by identifying every orbit to a single point in the reduced manifold, form the spaces of inhomogeneous

[2] The Kantowski–Sachs model is realized inside the horizon of the Schwarzschild solution. It is the only homogeneous model that cannot be obtained by imposing an extra isotropic symmetry on a Bianchi model.

models. For spherically symmetric space-times, for instance, orbits are 2-spheres and their orbit space on spatial slices, also called the *reduced manifold*, is the radial line.

To identify the fields of a reduced model for a given symmetry group, we need a complete basis for all invariant tensors involved. For the gravitational field in its canonical form, we decompose the spatial metric and the co-normal by using invariant 2-tensors and vectors on the symmetry orbits. Invariance conditions such as $s^*h_{ab} = h_{ab}$ are linear in the fields, and there are only finitely many independent equations if the symmetry group has a finite number of generators. In principle, the equations are not difficult to solve, but the amount of freedom left for solutions is not always easy to oversee and to order. We will discuss systematic classification methods in the context of connections, and for now assume that the invariance conditions have been solved and a basis of invariant tensor fields is known.

If, for certain label sets with indices $I \in \mathcal{I}$ and $K \in \mathcal{K}$, a basis of invariant 2-tensors is given by $\{h^I_{ab}\}_{I\in\mathcal{I}}$ and a basis of invariant co-vectors by $\{N^K_a\}_{K\in\mathcal{K}}$, then any invariant spatial metric can be written as $h_{ab} = \phi_I h^I_{ab}$ with fields ϕ_I on the reduced manifold, and the choice of frame, characterized by lapse and shift, gives rise to further fields N and N_I on the reduced manifold such that $N_a = N_I N^I_a$. While N is always a spatial scalar, the tensorial properties of the fields ϕ_I and N_I depend on the specific symmetry type. From the spatial metric and the co-normal, the general form $g_{ab} = -n_a n_b + h_{ab}$ of an invariant space-time metric is constructed.

Example 4.6 (Invariant metrics)
For homogeneity, the reduced manifold is a single point. All components ϕ_I are scalars and functions of time only, while $N_I = 0$.

For spherical symmetry, there is only one invariant vector field $\partial/\partial r$, and thus $N^a = M(\partial/\partial r)^a$. The spatial metric must be of the form $h_{ab} = \phi_1(\mathrm{d}r)_a(\mathrm{d}r)_b + \phi_2(\mathrm{d}\Omega^2)_{ab}$. Thus, a spherically symmetric space-time line element is of the general form

$$\mathrm{d}s^2 = -N^2\mathrm{d}t^2 + \phi_1(\mathrm{d}r + M\mathrm{d}t)^2 + \phi_2\mathrm{d}\Omega^2 \,. \tag{4.37}$$

All fields N, M, ϕ_1 and ϕ_2 depend only on r and t.

Once the general invariant form of the metric has been found, such as (4.37), inserting it into the full field equations provides the equations of motion for the reduced model. They form a set of second-order differential equations for the ϕ_I, depending on the frame via the lapse and shift components N and N_I.

4.2.1.2 Invariant connections

So far, except for examples, we have not specified the form of invariant tensors such as h^I_{ab}, or described any systematic scheme to derive a complete basis. Indeed, oftentimes the form of invariant metrics may appear rather obvious, as in the cases of homogeneous models, in which Lie theory helps us to classify all possible geometries, or for the familiar spherically symmetric space-times. It is nevertheless important to have a general classification scheme at hand, one that works for midi-superspace models, too, and which can show that, indeed, all possible invariant forms are covered for a given type of symmetry. Mathematically, this classification is much better developed for connections rather than metrics. The apparatus, based on principal fiber bundles, is quite involved, but it is an important tool in modern physics and thus is also useful for other purposes. From invariant connections, as we will see, one can find the form of invariant metrics. But the classification of invariant connections is already of interest for canonical gravity, since several of its most useful formulations are based on connection variables.

So far, we have encountered several examples of connections, primarily the Lorentz connection $\omega_a{}^I{}_J$. Since the parallel transport it generates leaves invariant the Minkowski metric η_{IJ}, the infinitesimal transport $v^a\omega_a{}^I{}_J \in \mathrm{so}(3,1)$ along an arbitrary vector field v^a takes values in the Lie algebra of the Lorentz group. In general, a connection behaves like a 1-form taking values in the Lie algebra of a so-called *structure group* G, but it is subject to non-tensorial transformation properties as seen for the Lorentz connection in Eq. (3.130). Choosing a basis $(\tau_i)^A{}_B$, we also denote a connection by its coefficients A_a^i in a decomposition $\omega_a{}^A{}_B = A_a^i(\tau_i)^A{}_B$, as used for the Ashtekar–Barbero connection: an su(2)-connection A_a^i which, using su(2)-generators $\tau_j = \frac{1}{2}i\sigma_j$ in terms of Pauli matrices σ_j, provides su(2)-connection 1-forms.

Example 4.7 (Homogeneous fields)
For an invariant connection, in addition to the structure group G we have a second group S, the symmetry group acting on spatial slices Σ whose elements are supposed to leave A_a^i invariant when applied to it by pull-back. If the whole Σ is a Lie group, it carries an obvious symmetry action by its own (left) multiplication; in this case we can identify the spatial manifold Σ with the symmetry group S as a manifold. The symmetry action is transitive, and we obtain the set of homogeneous Bianchi models. There is already a general sense of invariance: the left-invariant 1-forms $\boldsymbol{\omega}^I$ collected in the Maurer–Cartan form $\boldsymbol{\theta}_{\mathrm{MC}} = \boldsymbol{\omega}^I T_I$ *with generators T_I of the Lie algebra of S (see also the Appendix). For a matrix group, $\boldsymbol{\theta}_{\mathrm{MC}} = s^{-1}\mathrm{d}s$ can be computed directly from a parameterization of group elements s. From this expression, it is clear that $\boldsymbol{\theta}_{\mathrm{MC}}$ is invariant under left multiplication $L_r\colon s \mapsto rs$ by a constant $r \in S$ transforming the whole group manifold. The representation $s^{-1}(\mathrm{d}s)_a = \omega_a^I T_I$ provides useful means for computing the left-invariant 1-forms ω_a^I of a matrix Lie group. Every $\boldsymbol{\omega}^I$ obtained in this way is a 1-form on S such that $s^*\boldsymbol{\omega}^I = \boldsymbol{\omega}^I$ for all $s \in S$ acting on S by left multiplication, providing systematic means for the computation of invariant structures.*

As a whole, the 1-form $\boldsymbol{\theta}_{\mathrm{MC}}$ takes values in the Lie algebra of S. An invariant connection with structure group G obeys just the same invariance condition as $\boldsymbol{\theta}_{\mathrm{MC}}$, but it must take values in the Lie algebra of G, not S. A linear map is required, to be applied to the values of $\boldsymbol{\theta}_{\mathrm{MC}}$. There is no natural linear map $\tilde{\phi}\colon \mathcal{L}S \to \mathcal{L}G$ (not required to be an algebra homomorphism) between two arbitrary Lie algebras, $\mathcal{L}S$ and $\mathcal{L}G$. Any one of them gives rise to an invariant connection $A_a^i\tau_i\mathrm{d}x^a = \tilde{\phi}\circ\boldsymbol{\theta}_{\mathrm{MC}} = \tilde{\phi}_I^i\tau_i\boldsymbol{\omega}^I$, with components obtained from $\tilde{\phi}(T_I) =: \tilde{\phi}_I^i\tau_i$. This determines the set of all connections invariant under a transitive and freely acting symmetry group.

If the symmetry group still acts transitively but not freely, additional isotropies occur. At each point $x \in \Sigma$, there is a non-trivial isotropy subgroup $F_x < S$ whose elements leave x fixed. For different points, thanks to transitivity, the F_x only differ by conjugation (if $f(x) = x$ and $s(x) = y$, all transformations leaving y invariant are obtained as $sfs^{-1}(y) = y$), and we are allowed to pick one of them as a representative F. Now, we can identify the spatial manifold Σ (an orbit of the symmetry action) with the factor space S/F.

Example 4.8 (Euclidean group)
A group that combines translations and rotations in its action on $\mathbb{R}^3$ *is the Euclidean group* $S = \mathbb{R}^3 \rtimes \mathrm{SO}(3)$*, defined as the manifold* $\mathbb{R}^3 \times \mathrm{SO}(3)$ *of pairs* $t \in \mathbb{R}^3$ *and* $R \in \mathrm{SO}(3)$ *with the* semidirect product *rule*

$$(t_1, R_1)(t_2, R_2) = (t_1 + R_1 t_2, R_1 R_2)\,. \tag{4.38}$$

The Euclidean group acts on $\mathbb{R}^3$ *by* $(t, R)(x) = Rx + t$*. For any* $x \in \mathbb{R}^3$*, the isotropy subgroup is given by* $F_x = \{(\mathbb{I} - R)x, R) : R \in \mathrm{SO}(3)\} \cong \mathrm{SO}(3)$*. Symmetry orbits are* $S/F \cong \mathbb{R}^3 = \Sigma$*.*

In general, a factor space such as S/F is not a group, and so there is no Maurer–Cartan form defined on S/F to be used for invariant connections. Nevertheless, we can obtain invariant 1-forms if we choose an embedding $\iota\colon S/F \to S$ and take the pull-back $\iota^*\boldsymbol{\theta}_{\mathrm{MC}}$ of the Maurer–Cartan form defined on S. Normally, there is no unique embedding of this form, but different choices will simply amount to choosing coordinates on the homogeneous space. Using linear maps as before, invariant connections are then of the form $\tilde{\phi} \circ \iota^*\boldsymbol{\theta}_{\mathrm{MC}}$.

The pull back of the Maurer–Cartan form takes values in $\mathcal{L}S$. The map $\tilde{\phi}$ thus still maps from $\mathcal{L}S$ to $\mathcal{L}G$, even though the dimension of $\mathcal{L}S$ is now larger than that of Σ. There seem to be more components to an invariant connection, even though we are imposing stronger symmetries. The contradiction is resolved by considering the action of the isotropy subgroup F, which leaves points x in Σ fixed. It thus acts on the tangent space of x, linearly transforming the space on which $\tilde{\phi}$ is defined. No such transformation occurred in the free case, in which arbitrary $\tilde{\phi}$ gave rise to invariant connections. Now, there is an isotropy condition also for the map $\tilde{\phi}$, and the symmetry has not been imposed completely simply by using the left-invariant 1-forms. To handle isotropy conditions, additional tools to understand the gauge transformations, in particular principal fiber bundles, are useful. This will also allow us to derive the general form of invariant connections for groups not acting transitively, corresponding to midi-superspace models.

Connections We first have to extend the formal notion of a connection from a point-wise mapping $T_x\Sigma \to \mathcal{L}G$, as used for homogeneous configurations, to a global construct. This extension makes use of bundles, which like the tangent bundle $T\Sigma$ or the internal vector bundles attach a vector space of a given type, such as $\mathcal{L}G$, to each point of a manifold. A connection 1-form on a manifold M is then a map $A_a^i\colon T\Sigma \to E, v^a \mapsto A_a^i v^a$ from the tangent bundle $T\Sigma$ of Σ to a vector bundle E over the base manifold Σ. We thus handle the

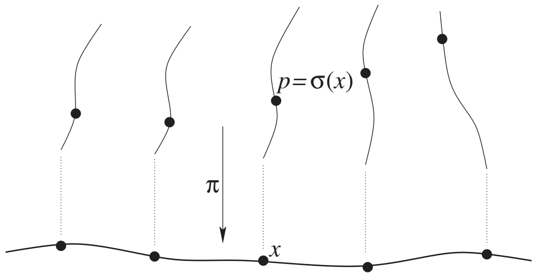

Fig. 4.3 A section σ of a fiber bundle assigns a unique element $p = \sigma(x) \in \pi^{-1}(x)$ to each element $x \in \Sigma$ of the base manifold.

point-wise mappings all at once, but this is not simply an obvious enlargement but requires careful definitions. In particular, the mapping as defined so far is not covariant owing to the transformation properties (3.130) of a connection 1-form.

A *vector bundle* E with vector space fiber V over the base manifold Σ is a manifold with a surjective projection map $\pi : E \to \Sigma$ such that for any local neighborhood U of Σ, $\pi^{-1}(U) \subset E$ is of the form $\pi^{-1}(U) = U \times V$. Locally, a vector bundle is simply the Cartesian product of its base manifold with the fiber, a property called *local triviality*. But globally, such a product decomposition may not be possible, for instance when twists exist as in the Moebius strip. (If the decomposition is possible globally, the vector bundle is called trivial.) Another example is the tangent space of the 2-sphere: the 2-sphere is not parallelizable, i.e. there is no global basis of nowhere-vanishing vector fields. If it were trivial, picking a basis of V would automatically provide a global basis of the bundle.

Given a vector bundle over Σ, there are two different notions of vector fields on Σ: tangent vectors v^a in $T\Sigma$, tangent to the base manifold; and internal vectors v^i in E, tangent to the fibers. Physically, vector bundles arise whenever we consider fields transforming under some group such as a gauge group. Sections of the bundle, i.e. maps $\sigma : \Sigma \to E$ such that $\pi \circ \sigma = \mathrm{id}$ (see Fig. 4.3), provide mathematical expressions for the fields.

A group G represented on the fiber V by a representation $\rho : G \times V \to V$ then determines how the entire field transforms. Also here, the tangent bundle is an example with vector fields transforming under the general linear group. The bundle notion allows us to formulate gauge transformations locally: for a trivial bundle $\Sigma \times V$ it may be sufficient to consider only global transformations of the action $\rho : G \times V \to V, (g, v^i) \mapsto \rho_g v^i = (\rho_g)^i_j v^j$ of G on V, transforming fields as $v^i(x) \mapsto \rho_g v^i(x)$ for all $g \in G$. For a non-trivial bundle, this is not possible, since there is no a-priori meaning of comparing vector fields or their gauge transformations over different neighborhoods of Σ. The only meaningful notion of gauge transformations is the local one: at each point x in the base manifold, there is an independent group element $g(x)$. The whole function $g : \Sigma \to G$ then locally transforms fields by $v^i(x) \mapsto \rho_{g(x)} v^i(x)$.

Just as we need to specify a covariant derivative for an unambiguous notion of differentiation on a curved manifold, we need a covariant derivative to determine how to differentiate fields in a vector bundle. Covariance requires that the transformation of every object agrees with the transformations of all the terms from which it is constructed: $\mathcal{D}_a v^i$ must be a contravariant internal vector (and a spatial covector), which requires a suitable transformation property for the covariant derivative operator (and the connection 1-forms it contains) as per

$$\rho_g(\mathcal{D}_a v^i) = (\rho_g)^i{}_j \mathcal{D}_a v^j = (\rho_g \mathcal{D}_a)(\rho_g v^i)\,. \tag{4.39}$$

Simple partial derivatives are not covariant under local gauge transformations,

$$\partial_a((\rho_g)^i{}_j v^j) = (\rho_g)^i{}_j(\partial_a v^j) + (\partial_a(\rho_g)^i{}_j)v^j \neq \rho_g(\partial_a v^i)\,. \tag{4.40}$$

A covariant derivative must have an additional contribution to cancel the second term, which is done by means of a connection 1-form A^i_a. It provides the covariant derivative $\mathcal{D}_a v^i := \partial_a v^i + A^j_a \tau^i_{jk} v^k$ where τ^i_{jk} is the representation matrix of the generator τ_j of the structure group G. (For instance for the adjoint action of G on its Lie algebra V, $\tau^i_{jk} = C^i_{jk}$ are the structure constants of the group.) Taking the derivative along a vector field t^a, we see that there is an extra term $\tau^i_j = t^a A^k_a \tau^i_{kj}$ which acts on internal vectors by the representation of the structure group. For $\mathcal{D}_a v^i$ to be covariant under local gauge transformations, the connection 1-form must transform as

$$\rho_g(A^k_a \tau^i_{kj}) = (\rho_g)^i_l (A^k_a \tau^l_{km})(\rho_{g^{-1}})^m_j - \partial_a(\rho_g)^i_m(\rho_{g^{-1}})^m_j \tag{4.41}$$

ensuring that $(\rho_g \mathcal{D}_a)(\rho_g v^i) = (\rho_g)^i_l(\partial_a v^l + A^k_a \tau^l_{km} v^m) = \rho_g(\mathcal{D}_a v^i)$ in combination with (4.40). Or, dropping the indices (i, j) and viewing $A^k_a \tau^i_{kj} =: (\mathbf{A})_a{}^i{}_j$ as a 1-form taking values in the transformation space of V,

$$\rho_g \mathbf{A} = \rho_g \mathbf{A} \rho_{g^{-1}} - (\mathrm{d}\rho_g)\rho_{g^{-1}}\,. \tag{4.42}$$

This defines the transformation properties of a connection 1-form, generalizing the transformation of a Lorentz connection in Eq. (3.130) (which was written for the inverse

$$\rho_{g^{-1}} \mathbf{A} = \rho_{g^{-1}} \mathbf{A} \rho_g + \rho_{g^{-1}}(\mathrm{d}\rho_g) \tag{4.43}$$

of the transformation considered here).

Principal fiber bundles Just as it is useful to view vector fields as global sections $v\colon \Sigma \to E$ of a vector bundle, rather than a collection of local assignments of vectors in the fiber V to all points in the base manifold Σ, there is a global notion for a connection locally represented by the connection 1-forms A^i_a on Σ. The required bundle notion is not a vector bundle but a principal fiber bundle over the same base manifold, Σ, with fibers not given by a vector space but by the structure group G itself. Locally, a principal fiber bundle P over the base manifold Σ with bundle projection $\pi\colon P \to \Sigma$ is thus of the form $\Sigma \times G$, but globally a decomposition is not possible in general. A principal fiber bundle carries an action of the group G by multiplication on its fibers (the right action of G on P), which is transitive. These can be taken as the defining properties of a *principal fiber bundle*.

A principal fiber bundle subsumes all possible gauges of a theory of connections. Choosing a gauge means that a unique point is specified on each fiber of the principal fiber bundle P, preferably in a smooth way. A gauge is thus described by a local section $\sigma : \mathcal{U} \to P$ such that $\pi(\sigma(x)) = x$ for all $x \in \mathcal{U} \subset \Sigma$ in an open neighborhood $\mathcal{U}$. Similarly, changing the gauge is accomplished by a section $h\colon \Sigma \to G$ acting on P by right multiplication, and changing the original section σ to σh (fiberwise product).

In a local trivialization, P can be written as the union of all $\mathcal{U} \times G$ where $\mathcal{U}$ runs through all neighborhoods in an open covering of Σ. Locally, we can thus map $P|_{\mathcal{U}} := \pi^{-1}(\mathcal{U}) \subset P$ to $\{(\pi(p), g(p))\}_{p\in\pi^{-1}(\mathcal{U})}$ with a map $g\colon P|_{\mathcal{U}} \to G$. The functions $g(p)$ can be used as local coordinates on the bundle, together with coordinates of Σ.

With these structures, we are able to reformulate the non-covariant term $g^{-1}(\mathrm{d}g)_a$ in the transformation of a connection as a covariant tensor on the bundle (rather than the base manifold). We use coordinates y^μ on P, x^i on Σ, and choose a section $\sigma : \mathcal{U} \to P$, $x^i \mapsto y^\mu(x)$ in coordinate form. In the non-covariant term, we write

$$g^{-1}\frac{\partial g}{\partial x^i}\mathrm{d}x^i = g^{-1}\frac{\partial g}{\partial y^\mu}\frac{\partial y^\mu}{\partial x^i}\mathrm{d}x^i = \sigma^*(g^{-1}\mathrm{d}g) = \sigma^*\boldsymbol{\theta}_{\mathrm{MC}}^{(P)}$$

with the *Maurer–Cartan form* $\boldsymbol{\theta}_{\mathrm{MC}}^{(P)} := g^{-1}\mathrm{d}g\colon TP \to \mathcal{L}G$. This form, defined on P, only varies along the fibers. On a single fiber, it is the Maurer–Cartan form $\boldsymbol{\theta}_{\mathrm{MC}}\colon TG \to \mathcal{L}G$ on the structure group, taking values in the Lie algebra.

Generalizing the Maurer–Cartan form of Lie groups, we define a *connection* on the principal G-fiber bundle P to be a 1-form $\boldsymbol{\theta}\colon TP \to \mathcal{L}G$ on P taking values in the Lie algebra $\mathcal{L}G$ of the structure group, such that $\boldsymbol{\theta}|_{TG} = \boldsymbol{\theta}_{\mathrm{MC}}$ is fixed along the fibers. The values along directions not tangent to fibers are not fixed, and correspond to the freedom contained in a local connection 1-form $\mathbf{A}$. With a local section $\sigma : \mathcal{U} \to P$ to specify a gauge, we have $\mathbf{A}_{(\sigma)} := \sigma^*\boldsymbol{\theta} = \sigma^{-1}\mathbf{A}_{(\mathbb{I})}\sigma + \sigma^{-1}\mathrm{d}\sigma$ where the *local connection 1-form* $\mathbf{A}_{(\sigma)}$ in the gauge σ is defined by this equation for a given $\boldsymbol{\theta}$, and $\sigma^{-1}\mathrm{d}\sigma$ arises from the pull-back of the fiber contribution to $\boldsymbol{\theta}$. The 1-form $\mathbf{A}_{(\mathbb{I})}$, used here for reference, is obtained for the "flat" gauge in which the section maps to the identity element everywhere. For a non-trivial principal fiber bundle, a gauge depends on the local trivialization chosen via the section, and does not exist globally.[3] (Accordingly, the gauge-dependent local connection 1-forms cannot always form a global object; only the P-connection $\boldsymbol{\theta}$ can.)

We change the gauge by right multiplication of sections $h\colon \sigma \mapsto \sigma h$, and a new local connection 1-form $\mathbf{A}_{(\sigma h)}$ is obtained from

$$\begin{aligned}\mathbf{A}_{(\sigma h)} &= (\sigma h)^*\boldsymbol{\theta} = (\sigma h)^{-1}\mathbf{A}_{(\mathbb{I})}\sigma h + (\sigma h^{-1})\mathrm{d}(\sigma h)\\ &= h^{-1}(\sigma^{-1}\mathbf{A}_{(\mathbb{I})}\sigma + \sigma^{-1}\mathrm{d}\sigma)h + h^{-1}\mathrm{d}h = h^{-1}\mathbf{A}_{(\sigma)}h + h^{-1}\mathrm{d}h\,.\end{aligned}$$

[3] Assume that there is a global section $\sigma\colon \Sigma \to P$. At each $x \in \Sigma$, there is for all $p \in \pi^{-1}(x)$ a unique $h_\sigma(p)$ such that $p = \sigma(x)\cdot h_\sigma(p)$ by right action. The map $P \to \Sigma \times G$, $p \mapsto (\pi(p), h_\sigma(p))$ gives a globally defined trivialization of P. Thus, a principal fiber bundle with a global section is trivial.

Thus, while $\boldsymbol{\theta}$ is simply a 1-form defined on the whole bundle, gauge transformations arise for its pull-backs along sections. In this way, principal fiber bundles allow a covariant, global formulation of the notion of connections by means of $\boldsymbol{\theta}$.

Classification of symmetric principal fiber bundles We now turn to the question of how to construct and classify all possible forms of connections on a space Σ, invariant under the action of a given symmetry group. Again, the invariance condition is given via the pull-back by elements of the symmetry group. However, if we work with a local connection 1-form, it can reasonably be required to be invariant only up to a gauge transformation: applying a symmetry transformation to the connection will produce the same physics even if the connection is not exactly invariant but only invariant up to gauge. The conditions to solve are now

$$s^*\mathbf{A} = \rho_{g(s)}\mathbf{A} = g(s)^{-1}\mathbf{A}g(s) + g(s)^{-1}\mathrm{d}g(s) \tag{4.44}$$

for a certain assignment of local gauge transformations $g(s)$ to the symmetry group elements. This complicates the situation: while the equations are still linear in A_a^i, one must know suitable $g(s)$ to solve them. In $g(s)$, however, the equations are not linear. Moreover, while applying a symmetry transformation twice leads to certain conditions for $g(s_1 s_2)$ in terms of $g(s_1)$ and $g(s_2)$, they do not distinguish $g\colon S \to G$ as a group homomorphism.

Example 4.9 (Spherically symmetric connections)
As derived by Cordero (1977), a general spherically symmetric SU(2)*-connection in polar coordinates, has the form*

$$\begin{aligned}\mathbf{A}(x,\vartheta,\varphi) = {} & A_1(x)\tau_3\mathrm{d}x + (A_2(x)\tau_1 + A_3(x)\tau_2)\mathrm{d}\vartheta \\ & + (A_2(x)\tau_2 - A_3(x)\tau_1)\sin\vartheta\,\mathrm{d}\varphi + \cos\vartheta\,\mathrm{d}\varphi\,\tau_3\end{aligned} \tag{4.45}$$

with three independent fields $A_I(x)$ depending only on the radial coordinate x, and with su(2)*-generators τ_i. (The gauge here is different from the one chosen by Cordero (1977).) An exactly invariant connection would only have the radial term $A_1(x)\tau_3\mathrm{d}x$, as expected for a spherically symmetric co-vector field which can only point radially. The extra terms do change when a rotation is applied, but this can be compensated for by a gauge transformation as shown by a direct calculation. Compared to the radial contribution, it is much more complicated to guess the general form of the angular ones. Here, a systematic classification scheme is most useful.*

The mathematical classification of invariant connections was developed by Kobayashi and Nomizu (1963, 1969); Brodbeck (1996) splits the problem of finding suitable $g(s)$ and solving for A_a^i into two steps, making use of principal fiber bundles. Then using the global notion of connections, an invariant connection $\boldsymbol{\alpha}$ on P is defined simply by the pull-back condition $s^*\boldsymbol{\alpha} = \boldsymbol{\alpha}$, while gauge transformations are implemented by evaluating $\boldsymbol{\alpha}$ at different points along the fibers. Gauge transformations and symmetry actions do not mix as they did in (4.44). However, to evaluate the condition of symmetry, we now need an action of the symmetry group on P, the space on which $\boldsymbol{\alpha}$ is defined, while we are initially given an action on the base manifold Σ only. Lifting the action to P amounts to finding suitable $g(s)$, and so we are not completely freed from that task. The great advantage now is that an extension to P can be analyzed without using connections or the invariance condition.

Let us start with a given action of the symmetry group S on Σ. In general, for a point $x \in \Sigma$ there may be an isotropy subgroup $F_x < S$ with $f(x) = x$ for all $f \in F_x$. Lifting the action to automorphisms on a principal fiber bundle $\pi: P \to \Sigma$ over Σ with structure group G must produce an action $F_x: \pi^{-1}(x) \to \pi^{-1}(x)$ on the fiber over x, commuting with the right action of G on the bundle. The action from fiber to fiber is determined by the action of S on Σ. But while F_x acts trivially on Σ, its action on $\pi^{-1}(x)$ may be non-trivial. Here, the different lifts to P are realized. If we can determine all possible actions of F along the fibers compatible with the action of S on a principal fiber bundle, the classification of symmetric principal fiber bundles is complete.

Let us assume that an action of F on the fibers is given. Since the right action of the bundle is transitive, to each point $p \in \pi^{-1}(x)$ on the fiber is uniquely assigned a group homomorphism $\lambda_p: F \to G$ by $f(p) =: p \cdot \lambda_p(f)$ for all $f \in F$. To verify that λ_p is indeed a group homomorphism, we first derive its behavior under gauge transformations, changing the point p along the fiber to $p' = p \cdot g$ with $g \in G$:

$$p' \cdot \lambda_{p'}(f) = f(p') = f(p \cdot g) = f(p) \cdot g = (p \cdot \lambda_p(f)) \cdot g = p' \cdot \mathrm{Ad}_{g^{-1}} \lambda_p(f)$$

with the adjoint action $\mathrm{Ad}_g h = ghg^{-1}$ of G on itself. (By the automorphism condition above, the action of f commutes with right multiplication by g.) Thus, $\lambda_{p \cdot g} = \mathrm{Ad}_{g^{-1}} \circ \lambda_p$. Then, we have

$$(f_1 \circ f_2)(p) = f_1(p \cdot \lambda_p(f_2)) = (p \cdot \lambda_p(f_2)) \cdot \mathrm{Ad}_{\lambda_p(f_2)^{-1}} \lambda_p(f_1) = p \cdot (\lambda_p(f_1) \cdot \lambda_p(f_2))$$

and the homomorphism property follows. To summarize, the action on fibers is characterized by a map $\lambda: P \times F \to G, (p, f) \mapsto \lambda_p(f)$ which obeys the relation $\lambda_{p \cdot g} = \mathrm{Ad}_{g^{-1}} \circ \lambda_p$. An important element of the classification of symmetric principal fiber bundles is thus an equivalence class $[\lambda]$ of group homomorphisms $\lambda: F \to G$ up to conjugation. Changing λ within a conjugacy class simply amounts to a change of gauge.

The remaining elements of the classification amount to the structure of a reduced bundle. We expect that the base manifold, which is Σ for P, can be reduced to the orbit space Σ/S, since the structure of a symmetric bundle is not supposed to change when applying S. The reduced bundle is indeed a subbundle of the restricted bundle $P|_{\Sigma/S}$. It turns out that the structure group must also be reduced. For an explicit construction of the bundle (or later on invariant connections on it) we will be working with a specific homomorphism λ, not with a conjugacy class. Gauge transformations on P then split into two different classes: those that fix the conjugacy class (and thus the image $\lambda(F)$ in G) and which will be represented as gauge transformations of the reduced bundle, and those that would change the conjugacy class and thus the reduced bundle. Reduced gauge transformations in the former class form the centralizer $Z_G(\lambda(F))$ of $\lambda(F)$ in G, i.e. the set of all G-elements commuting with all elements of $\lambda(F)$. This determines the structure of the reduced bundle

$$Q_\lambda(B, Z_G(\lambda(G)), \pi_Q) := \{p \in P_{|B} : \lambda_p = \lambda\} \tag{4.46}$$

for a given homomorphism $\lambda\colon F \to G$. Conjugation by G on $\lambda(F)$ implies a corresponding action on the set of reduced bundles. In summary, following Brodbeck (1996), we have

Theorem 4.1 *An S-symmetric principal fiber bundle $P(\Sigma, G, \pi)$ with isotropy subgroup $F \leq S$ of the action of S on Σ is uniquely characterized by a conjugacy class $[(\lambda, Q_\lambda)]$ of homomorphisms $\lambda\colon F \to G$ together with a reduced bundle $Q_\lambda(\Sigma/S, Z_G(\lambda(F)), \pi_Q)$.*

The action of a symmetry group often has different types of isotropy subgroup depending on the point $x \in \Sigma$. For spherical symmetry, for instance, $S = \mathrm{SO}(3)$ and $F = \mathrm{SO}(2)$ outside the center while $F = \mathrm{SO}(3)$ at the center. In those cases, the bundle must be split by patching regions of different isotropy types.

In order to determine all conjugacy classes of homomorphisms $\lambda\colon F \to G$ we can make use of the relation

$$\mathrm{Hom}(F, G)/\mathrm{Ad} \cong \mathrm{Hom}(F, T(G))/W(G) \tag{4.47}$$

where $T(G)$ is a maximal torus and $W(G)$ the Weyl group of G.[4] Physically, different conjugacy classes correspond to different sectors of the gauge theory, or different values of a topological charge. In spherically symmetric electromagnetism, this is the well-known magnetic charge; see Example 4.10. Equation (4.47) will also be applied in Example 4.13.

Classification of invariant connections Given a symmetric principal fiber bundle P classified by $[(\lambda, Q_\lambda)]$, the invariant connections it allows must satisfy $s^*\boldsymbol{\alpha} = \boldsymbol{\alpha}$ for all $s \in S$. Since Q is a subbundle of P, we can restrict the connection to obtain a connection $\tilde{\boldsymbol{\alpha}}$ on the reduced bundle Q. This is not completely obvious, since a connection on Q must take values in the Lie algebra of the reduced structure group, while the restriction of a connection on P is initially expected to take values in the Lie algebra of G. An S-invariant connection on P, on the other hand, automatically takes values in the Lie algebra of the reduced structure group $Z_G(\lambda(F))$: at an arbitrary point $p \in P$ we choose a vector v in T_pP such that $\pi_* v \in \sigma_* T_{\pi(p)}(\Sigma/S)$ where σ is the embedding of Σ/S into Σ. The condition means that v does not have components along symmetry orbits, and is thus fixed by the action of the isotropy group: $\mathrm{d}f(v) = v$. (Any such vector v is along $\partial/\partial r$ in spherical symmetry.) By definition, the pull-back of $\boldsymbol{\alpha}$ by $f \in F$, applied to v, is $f^\star\boldsymbol{\alpha}_p(v) = \boldsymbol{\alpha}_{f(p)}(\mathrm{d}f(v))$ which now equals $\boldsymbol{\alpha}_{f(p)}(v)$. On the other hand, as used in the classification of symmetric bundles, f acts as a gauge transformation by $\lambda_p(f)$, under which $\boldsymbol{\alpha}$ transforms as $\boldsymbol{\alpha}_{f(p)}(v) = \mathrm{Ad}_{\lambda_p(f)^{-1}}\boldsymbol{\alpha}_p(v)$. We have assumed the connection $\boldsymbol{\alpha}$ to be S-invariant, thus $f^\star\boldsymbol{\alpha}_p(v) = \mathrm{Ad}_{\lambda_p(f)^{-1}}\boldsymbol{\alpha}_p(v) = \boldsymbol{\alpha}_p(v)$ for all $f \in F$ such that

[4] The maximal torus T of a compact Lie group G is a maximal compact, connected, Abelian subgroup. It can be shown that it non-trivially intersects every conjugacy class in the group, but sometimes conjugacy classes are intersected more than once. The Weyl group W (defined as the normalizer of T modulo its centralizer in G) then acts transitively on $T \cap [g]$ for an arbitrary $g \in G$.

$\alpha_p(v) \in \mathcal{L}Z_G(\lambda_p(F))$. An invariant connection $\boldsymbol{\alpha}$ on P can thus be restricted to a connection on the bundle Q_λ with structure group Z_λ. Such a connection on the reduced bundle is the first part of the classification of invariant connections.

The reduced connection is not the complete information contained in an invariant one on P. (In the example of spherical symmetry, this connection would just be the radial part.) From $\boldsymbol{\alpha}$ we further construct a linear map $\Lambda_p \colon \mathcal{L}S \to \mathcal{L}G$, $X \mapsto \boldsymbol{\alpha}_p(\tilde{X})$ for any $p \in P$, evaluating $\boldsymbol{\alpha}$ on the vector field $\tilde{X}$ on P, given by $\tilde{X}\psi := \mathrm{d}(\exp(tX)^\star\psi)/\mathrm{d}t|_{t=0}$ for $X \in \mathcal{L}S$ and ψ a differentiable function on P. If $X \in \mathcal{L}F$, $\tilde{X}$ is vertical; we have $\Lambda_p(X) = \mathrm{d}\lambda_p(X)$ in terms of λ_p already defined for the symmetric bundle, irrespective of the connection. By gauge choice, λ can be held constant along Σ/S and so Λ_p evaluated on the Lie algebra of the isotropy subgroup does not provide free fields. But the remaining components, $\Lambda_p|_{\mathcal{L}F_\perp}$ making use of a decomposition $\mathcal{L}S = \mathcal{L}F \oplus \mathcal{L}F_\perp$, for instance with respect to the Killing metric for a semisimple Lie algebra, do yield independent information about the invariant connection $\boldsymbol{\alpha}$, subject only to the condition

$$\Lambda_p(\mathrm{Ad}_f(X)) = \mathrm{Ad}_{\lambda_p(f)}(\Lambda_p(X)) \quad \text{for} \quad f \in F, X \in \mathcal{L}S\,. \tag{4.48}$$

This equation, following from the transformation of $\boldsymbol{\alpha}$ under the adjoint representation, provides a set of equations to determine the form of the components Λ, and thus of invariant connections.

In addition to the reduced connection $\tilde{\boldsymbol{\alpha}}$, we have the components of a field $\tilde{\phi} \colon Q \to \mathcal{L}G \otimes \mathcal{L}F_\perp^\star$ determined by $\Lambda_p|_{\mathcal{L}F_\perp}$. This field can be regarded as having $\dim \mathcal{L}F_\perp$ components of $\mathcal{L}G$-valued scalar fields. We have now completed the characterization of invariant connections, and conclude with Brodbeck (1996):

Theorem 4.2 (Generalized Wang theorem) *Let $P(\Sigma, G)$ be an S-symmetric principal fiber bundle classified by $[(\lambda, Q_\lambda)]$ according to Theorem 4.1, and let $\boldsymbol{\alpha}$ be an S-invariant connection on P.*

Then, in a gauge given by (λ, Q_λ) the connection $\boldsymbol{\alpha}$ is uniquely determined by a reduced connection $\tilde{\boldsymbol{\alpha}}$ *on Q_λ and a* scalar field $\tilde{\phi} \colon Q_\lambda \times \mathcal{L}F_\perp \to \mathcal{L}G$ *obeying Eq. (4.48).*

As components of an invariant P-connection, $\tilde{\phi}$ transforms under G. However, if we solve Eq. (4.48), we must use a specific selection for the isotropy subgroup F (and thus fix its conjugacy class in G). Not all G-elements will leave this choice and thus the solution space of (4.48) for a given F invariant. Again, it is only elements of $Z_G(\lambda(F))$, the reduced structure group, that do so; they form the group under which the scalar fields transform.

From a reduced connection $\tilde{\boldsymbol{\alpha}}$ on Q_λ and scalars $\tilde{\phi}$ solving (4.48) an invariant connection α on P can be reconstructed, completing the classification. With the (local) decomposition $\Sigma \cong \Sigma/S \times S/F$, we have

$$\boldsymbol{\alpha} = \tilde{\boldsymbol{\alpha}} + \boldsymbol{\alpha}_{S/F} \tag{4.49}$$

where $\boldsymbol{\alpha}_{S/F}$ is given by the scalars, or rather, the whole Λ, via $\Lambda \circ \iota^\star\boldsymbol{\theta}_{\mathrm{MC}}$. Here, $\boldsymbol{\theta}_{\mathrm{MC}}$ is the Maurer–Cartan form on S taking values in $\mathcal{L}S$, and $\iota \colon S/F \hookrightarrow S$ is a (local) embedding. Choosing ι contributes to the gauge of the reconstructed connection.

Example 4.10 (Spherically symmetric electromagnetism)
With $S = \mathrm{SU}(2)$, $F = \mathrm{U}(1)$ *(which we realize as a subgroup of* $\mathrm{SU}(2)$ *by* $F = \exp\langle\tau_3\rangle$*), and* $G = \mathrm{U}(1)$ *we have conjugacy classes of homomorphism represented by* $\lambda_n \colon \mathrm{U}(1) \to \mathrm{U}(1)$, $\exp(\varphi\tau_3) \mapsto \exp(in\varphi)$ *for integer* n*. For an Abelian structure group,* $\mathrm{Ad}_{\lambda_n(f)} = \mathrm{id}$ *and the condition (4.48) reads* $\Lambda_p(\mathrm{Ad}_f X) = \Lambda_p(X)$*. The adjoint action of* F *on* $\mathcal{L}S$ *leaves only the subalgebra* $\mathcal{L}F$ *invariant, and so (4.48) implies that* $\Lambda|_{\mathcal{L}F_\perp} = 0$*, while* $\Lambda|_{\mathcal{L}F} = \mathrm{d}\lambda_n$ *is, as always, fully determined by the homomorphism* λ_n*. For Abelian structure groups, no free scalar fields arise in the classification of invariant connections.*

The only free part in a spherically symmetric U(1)-connection is thus the reduced connection $\tilde{\mathbf{A}} = A(r)\mathrm{d}r$*. But there is an extra contribution from* $\Lambda|_{\mathcal{L}F}$*, whose form follows from* $\mathbf{A}_{S/F}$ *for the homogeneous space of a sphere. Together with results of Example 4.14 and Exercise 4.8, the map* Λ *realized here implies that* $\mathbf{A}_{S/F} = \mathrm{d}\lambda_n(\tau_3)\cos\vartheta\,\mathrm{d}\varphi = in\cos\vartheta\,\mathrm{d}\varphi$*. If we remove the generator* i *of* $\mathcal{L}\mathrm{U}(1)$*, a connection of the form* $\mathbf{A}_{S/F}$ *corresponds to a field strength* $\mathbf{F}_{S/F} = -n\sin\vartheta\,\mathrm{d}\vartheta \wedge \mathrm{d}\varphi$*, and a (dedensitized) radial magnetic field of magnitude* $B^r = -n(r^2\sin\vartheta)^{-1}\sin\vartheta = -n/r^2$*. This is the magnetic field of a magnetic monopole of charge* n*, providing a physical example for the topological-charge parameter* n *classifying inequivalent symmetric bundles via* λ_n*. (Notice that the language of principal fiber bundles has allowed us to express the field of a magnetic monopole in a well-defined way, avoiding the string-like singularities in distributional realizations related to Dirac monopoles.)*

Additional examples for applications of this general procedure will be provided after a discussion of how invariant connections can be used to derive invariant metrics.

4.2.1.3 Canonical reduction

In many formulations of gravity, connections appear as one half of a set of canonical variables. Their classification is now available, but we also have to know how their momenta can be reduced to symmetric forms. A reduction must obviously occur, for the symmetry has diminished the number of independent components of connections. A symplectic phase space then requires the same number of conditions for their conjugate momenta.

Tensorial objects other than connections do not have the same degree of theoretical classification for their invariant forms. Fortunately, if the reduction is known for one canonical half, that of the other can be read off much more simply, using duality. If we know a basis of all invariant connections A_a^I (for instance, the left-invariant 1-forms for homogeneous models) for a given symmetry type, their momenta must be built from dual invariant fields X_I^a satisfying $A_a^I X_J^a = \delta_J^I$. This condition ensures that the same number of independent fields arises for configuration variables and momenta. Moreover, expansion coefficients by the invariant basis are automatically canonically conjugate: general invariant fields $A_a^i = \tilde{\phi}_I^i A_a^I$ and $E_i^a = \tilde{X}\tilde{p}_i^I X_I^a$, expanded in suitable invariant bases and with an invariant density $\tilde{X}$ (for instance, $\tilde{X} = |\det X_I^a|^{-1}$) to provide the density weight on symmetry orbits, satisfy $\int \mathrm{d}^3x\,\dot{A}_a^i E_i^a = \int \mathrm{d}^3x\,\tilde{X}\dot{\tilde{\phi}}_I^i \tilde{p}_i^I$. Once trivial integrations over symmetric orbits are performed, $\tilde{\phi}_I^i$ and $\tilde{p}_i^I$ appear as canonically conjugate fields; see also the examples in the next section.

Example 4.11 (Reduced densities)
Homogeneity is realized with respect to left-invariant 1-forms ω_a^I *and dual vector fields* X_I^a*, providing an invariant density* $\tilde{X} = |\det X_I^a|^{-1} = |\det \omega_a^I|$*. The symplectic structure follows from* $\int \mathrm{d}^3x \tilde{X} \dot{\tilde{\phi}}_I^i \tilde{p}_i^I = V_0 \dot{\tilde{\phi}}_I^i \tilde{p}_i^I$ *with* $V_0 = \int \mathrm{d}^3x \tilde{X}$ *integrated over some finite region. The homogeneous variables* $\tilde{\phi}_I^i$ *and* $\tilde{p}_i^I$ *have lost the density weight of the original fields.*

In spherical symmetry, some variables remain as 1-dimensional densities on the radial line. Factors of $\tilde{X} = \sin\vartheta$ *provide the density weight on spherical orbits.*

In particular, the classification of invariant connections can be used to tell us what the general invariant form of metric tensors is. As we have seen, in the canonical formulations of gravity in terms of connection variables, conjugate momentum variables E_i^a can be given the geometrical interpretation of densitized triads. They, in turn, directly lead to the spatial metric $h^{ab} = E_i^a E^{bi}/|\det(E_j^c)|$ whose invariant form then follows from that of the densitized triad. Momentum variables conjugate to the metric in ADM formulations, finally, are obtained by duality as in the case of connection variables. At this stage, we have a systematic procedure to determine the complete form of all invariant tensor fields as they occur in canonical theories of gravity.

The relation to metric variables provides one further condition on the allowed forms: non-degeneracy of the metric. If we start with a class of invariant connections, we are led to invariant densitized triads, and then to metrics. Not all of the possible sectors of invariant connections, classified by topological charges, may allow non-degenerate metrics. Thus, not all values for the charge, which might be allowed for a non-gravitational gauge theory, will be possible. We will see such an example in spherical symmetry.

4.2.1.4 Examples

We can now complete the examples already introduced in the course of this section.

Example 4.12 (Homogeneous models, once more)
In Bianchi models the transitive symmetry group acts freely on Σ*, which we locally identify with the group manifold* S*. We have three generators* T_I*,* $1 \leq I \leq 3$*, of* $\mathcal{L}S$*, with relations* $[T_I, T_J] = C^K_{IJ} T_K$*. By definition, the structure constants* C^K_{IJ} *of* $\mathcal{L}S$ *fulfill* $C^J_{IJ} = 0$ *for class A models. From the Maurer–Cartan form* $\boldsymbol{\theta}_{\mathrm{MC}} = s^{-1}\mathrm{d}s = \boldsymbol{\omega}^I T_I$ *on* S *we obtain left-invariant 1-forms* $\boldsymbol{\omega}^I$ *on* S*, fulfilling the Maurer–Cartan equations*

$$\mathrm{d}\boldsymbol{\omega}^I = -\tfrac{1}{2} C^I_{JK} \boldsymbol{\omega}^J \wedge \boldsymbol{\omega}^K \,. \tag{4.50}$$

The reduced manifold Σ/S *is a single point for homogeneous models, and any reduced bundle is trivial. Due to* $F = \{1\}$ *for a freely acting group, there is only one homomorphism* $\lambda\colon F \to G, 1 \mapsto 1$*. Only one type of symmetric bundle exists, with structure group* $Z_G(\lambda(F)) = G$*.*

With the reduced manifold a single point, there is no reduced connection $\tilde{\boldsymbol{\alpha}}$ *contributing to an invariant one. Only the scalar fields* $\tilde{\phi}\colon \mathcal{L}S \to \mathcal{L}G, T_I \mapsto \tilde{\phi}(T_I) =: \tilde{\phi}_I^i \tau_i$ *appear, unrestricted by (4.48) which is empty in this case. To reconstruct an invariant connection,* *we use the obvious embedding* $\iota = \mathrm{id}\colon S/F \equiv S \hookrightarrow S$. *Invariant connection 1-forms then take the form* $\mathbf{A} = \tilde{\phi} \circ \boldsymbol{\theta}_{\mathrm{MC}} = \tilde{\phi}^i_I \tau_i \boldsymbol{\omega}^I = A^i_a \tau_i \mathrm{d}x^a$ *with matrices* τ_i *generating* $\mathcal{L}G$.

Left-invariant vector fields X^a_I *on the Lie algebra, obeying* $\boldsymbol{\omega}^I(X_J) = \delta^I_J$, *provide the dual form of invariant vector fields. Analogously to the Maurer–Cartan relations for left-invariant 1-forms, they satisfy the Lie brackets* $[X_I, X_J] = C^K_{IJ} X_K$. *Momenta canonically conjugate to* $A^i_a = \tilde{\phi}^i_I \omega^I_a$ *can be written as* $E^a_i = \sqrt{h_0}\, \tilde{p}^I_i X^a_I$ *with independent components* $\tilde{p}^I_i$. *In addition to* X^a_I, *we use* $h_0 = \det(\omega^I_a)^2$ *as the determinant of the left-invariant metric* $h_{0\,ab} := \sum_I \omega^I_a \omega^I_b$ *on* Σ, *providing the density weight of* E^a_i. *To verify that the* $\tilde{p}^I_i$ *are canonically conjugate to the* $\tilde{\phi}^i_I$, *we compute*

$$\frac{1}{8\pi\gamma G}\int \mathrm{d}^3x\, \dot{A}^i_a E^a_i = \frac{1}{8\pi\gamma G}\int \mathrm{d}^3x\, \sqrt{h_0}\, \dot{\tilde{\phi}}^i_I \tilde{p}^J_i \omega^I_a X^a_J = \frac{V_0}{8\pi\gamma G} \dot{\tilde{\phi}}^i_I \tilde{p}^I_i\,,$$

and obtain

$$\{\tilde{\phi}^i_I, \tilde{p}^J_j\} = 8\pi\gamma G V_0 \delta^i_j \delta^J_I \tag{4.51}$$

with the volume $V_0 := \int \mathrm{d}^3x \sqrt{h_0}$ *of* Σ *measured with the invariant metric* h_0. *(If* Σ *is non-compact, we choose a finite region which we over integrate. Thanks to the homogeneity, results from the model will be insensitive to the choice.)*

This completes the invariant structure of canonical fields for homogeneous models without a non-trivial isotropy subgroup, and verifies what we already saw in Example 4.7 before the systematic classification. The general discussion will be most important when isotropy subgroups enter the game.

Example 4.13 (Isotropic models)
For a Bianchi model with additional isotropies, including FLRW models, the symmetry group is a semidirect product $S = N \rtimes_\rho F$ *of the isotropy subgroup* F *and the translational subgroup* N, *one of the Bianchi groups. (See Example 4.8 for the specific case of the Euclidean group.) The isotropy subgroup may be* SO(3) *for fully isotropic models, or* U(1) *for models with a single isotropic axis (called locally rotationally symmetric, LRS, as introduced by Ellis and MacCallum (1969)).*

Composition in this group of semidirect-product form is defined by

$$(n_1, f_1)(n_2, f_2) := (n_1 \rho(f_1)(n_2), f_1 f_2) \tag{4.52}$$

with a group homomorphism $\rho\colon F \to \mathrm{Aut}N$ *into the automorphism group of* N *(that is,* $\rho(f)\colon N \to N$ *is an isomorphism for all* $f \in F$*). Inverse elements are* $(n, f)^{-1} = (\rho(f^{-1})(n^{-1}), f^{-1})$. *To determine the form of invariant connections, we compute the Maurer–Cartan form on* S

$$\begin{aligned}
\boldsymbol{\theta}^{(S)}_{\mathrm{MC}}(n, f) &= (n, f)^{-1}\mathrm{d}(n, f) = (\rho(f^{-1})(n^{-1}), f^{-1})(\mathrm{d}n, \mathrm{d}f)\\
&= (\rho(f^{-1})(n^{-1})\rho(f^{-1})(\mathrm{d}n), f^{-1}\mathrm{d}f) = (\rho(f^{-1})(n^{-1}\mathrm{d}n), f^{-1}\mathrm{d}f)\\
&= \left(\rho(f^{-1})(\boldsymbol{\theta}^{(N)}_{\mathrm{MC}}(n)), \boldsymbol{\theta}^{(F)}_{\mathrm{MC}}(f)\right)\,.
\end{aligned} \tag{4.53}$$

in terms of $\boldsymbol{\theta}_{\mathrm{MC}}^{(N)}$ *on* N *and* $\boldsymbol{\theta}_{\mathrm{MC}}^{(F)}$ *on* F*. If we choose the embedding* $\iota\colon S/F = N \hookrightarrow S$ *as* $\iota\colon n \mapsto (n, 1)$*, we have* $\iota^*\boldsymbol{\theta}_{\mathrm{MC}}^{(S)} = \boldsymbol{\theta}_{\mathrm{MC}}^{(N)}$*. A reconstructed connection takes the same form* $\tilde{\phi} \circ \iota^*\boldsymbol{\theta}_{\mathrm{MC}}^{(S)} = \tilde{\phi}_I^i \boldsymbol{\omega}^I \tau_i$ *as seen for anisotropic models before. (Now,* $\boldsymbol{\omega}^I$ *are left-invariant 1-forms on the translation group* N*, which earlier was the full symmetry group.) However, here* $\tilde{\phi}$ *is constrained by Eq. (4.48), and only a subset of linear maps* $\tilde{\phi}$ *from* $\mathcal{L}N$*, generated by* T_I*, to* $\mathcal{L}G$ *provides isotropic connections.*

Solving Eq. (4.48) requires us to treat LRS and isotropic models separately. To be specific, we specialize the structure group G *to* $\mathrm{SU}(2)$ *as realized for the Ashtekar–Barbero connection. In the first case (LRS) we choose* $\mathcal{L}F = \langle \tau_3 \rangle$*, whereas in the second case (isotropy) we have* $\mathcal{L}F = \langle \tau_1, \tau_2, \tau_3 \rangle$ *(denoting by* $\langle \cdot \rangle$ *the linear span). Equation (4.48) can be written infinitesimally as*

$$\tilde{\phi}(\mathrm{ad}_{\tau_i}(T_I)) = \mathrm{ad}_{\mathrm{d}\lambda(\tau_i)} \tilde{\phi}(T_I) = [\mathrm{d}\lambda(\tau_i), \tilde{\phi}(T_I)]$$

($i = 3$ *for LRS,* $1 \leq i \leq 3$ *for isotropy). On* $\mathcal{L}N = \mathcal{L}F_\perp$ *the isotropy subgroup* F *acts by rotation,* $\mathrm{ad}_{\tau_i}(T_I) = \epsilon_{iIK} T_K$*. This is expected for an additional rotational symmetry, but it also follows automatically from the derivative of the representation* ρ *defining the semidirect product* S*: conjugation on the left-hand side of (4.48) is* $\mathrm{Ad}_{(1,f)}(n, 1) = (1, f)(n, 1)(1, f^{-1}) = (\rho(f)(n), 1)$ *according to the composition in* S*.*

To evaluate (4.48) further, we determine the possible conjugacy classes of homomorphism $\lambda\colon F = \mathrm{U}(1) \to G = \mathrm{SU}(2)$*. Applying Eq. (4.47), we use the standard maximal torus*

$$T(\mathrm{SU}(2)) = \{\mathrm{diag}(z, z^{-1}) : z \in \mathrm{U}(1)\} \cong \mathrm{U}(1)$$

of $\mathrm{SU}(2)$*, with the Weyl group of* $\mathrm{SU}(2)$ *given by the permutation group of two elements,* $W(\mathrm{SU}(2)) \cong S_2$*, its generator acting on* $T(\mathrm{SU}(2))$ *by* $\mathrm{diag}(z, z^{-1}) \mapsto \mathrm{diag}(z^{-1}, z)$*. Up to conjugation, all homomorphisms in* $\mathrm{Hom}(\mathrm{U}(1), T(\mathrm{SU}(2)))$ *are given by*

$$\lambda_k\colon z \mapsto \mathrm{diag}(z^k, z^{-k}) \tag{4.54}$$

for $k \in \mathbb{Z}$*, factored out by the action of the Weyl group to leave only the maps* λ_k*,* $k \in \mathbb{N}_0$*, as representatives of all conjugacy classes of homomorphism.*

For LRS models, we choose representatives

$$\lambda_k\colon \mathrm{U}(1) \to \mathrm{SU}(2), \quad \exp t\tau_3 \mapsto \exp kt\tau_3$$

for $k \in \mathbb{N}_0 = \{0, 1, \ldots\}$*. For the components* $\tilde{\phi}_I^i$*, Eq. (4.48) takes the form* $\epsilon_{3IK}\tilde{\phi}_K^j = k\epsilon_{3lj}\tilde{\phi}_I^l$*. There is a non-trivial solution only if* $k = 1$*, in which case* $\tilde{\phi}$ *can be written as* $\tilde{\phi}_1 = \tilde{a}\tau_1 + \tilde{b}\tau_2$*,* $\tilde{\phi}_2 = -\tilde{b}\tau_1 + \tilde{a}\tau_2$*,* $\tilde{\phi}_3 = \tilde{c}\tau_3$ *with arbitrary numbers* $\tilde{a}$*,* $\tilde{b}$*,* $\tilde{c}$*. Their conjugate momenta take the form* $\tilde{p}^1 = \frac{1}{2}(\tilde{p}_a\tau_1 + \tilde{p}_b\tau_2)$*,* $\tilde{p}^2 = \frac{1}{2}(-\tilde{p}_b\tau_1 + \tilde{p}_a\tau_2)$*,* $\tilde{p}^3 = \tilde{p}_c\tau_3$*, with non-vanishing Poisson brackets* $\{\tilde{a}, \tilde{p}_a\} = \{\tilde{b}, \tilde{p}_b\} = \{\tilde{c}, \tilde{p}_c\} = 8\pi\gamma G V_0$*. There is remaining gauge freedom from the reduced structure group* $Z_\lambda \cong \mathrm{U}(1)$ *which rotates the pairs* $(\tilde{a}, \tilde{b})$ *and* $(\tilde{p}_a, \tilde{p}_b)$*. The gauge invariant combinations are only* $\sqrt{\tilde{a}^2 + \tilde{b}^2}$ *and its momentum* $(\tilde{a}\tilde{p}_a + \tilde{b}\tilde{p}_b)/\sqrt{\tilde{a}^2 + \tilde{b}^2}$*.*

In the case of isotropic models with $S = \mathrm{SU}(2)$*, we have only two homomorphisms up to conjugation:* $\lambda_0\colon \mathrm{SU}(2) \to \mathrm{SU}(2)$*,* $f \mapsto 1$ *and* $\lambda_1 = \mathrm{id}$*. For* λ_0*, (4.48) in the form*

$\epsilon_{iIK}\tilde{\phi}^j_K = 0$ lacks non-trivial solutions. For λ_1, we solve $\epsilon_{iIK}\tilde{\phi}^j_K = \epsilon_{ilj}\tilde{\phi}^l_I$ as we did for LRS models with $k = 1$, resulting in $\tilde{\phi}^i_I = \tilde{c}\delta^i_I$ with an arbitrary $\tilde{c}$. In this case the conjugate momenta can be written as $\tilde{p}^I_i = \tilde{p}\delta^I_i$, and from $V_0(8\pi\gamma G)^{-1}\dot{\tilde{\phi}}^i_I\tilde{p}^I_i = 3V_0(8\pi\gamma G)^{-1}\dot{\tilde{c}}\tilde{p}$ we have the Poisson bracket $\{\tilde{c}, \tilde{p}\} = \frac{8\pi}{3}G\gamma V_0$. Obviously, $\tilde{p}$ must be related to the scale factor of an isotropic metric. Indeed, computing the isotropic reduction of a Bianchi IX metric following from the left-invariant 1-forms of SU(2)*, one obtains a closed FLRW metric with scale factor $a = 2\tilde{a} = 2\sqrt{|\tilde{p}|}$; see Exercise 4.4. (The Bianchi I model does not allow fixing the numerical relationship, since the scale factor of its isotropic reduction can be rescaled arbitrarily.)*

In both classes of isotropic submodel, there is a unique non-trivial sector and no topological charge appears.

For an example with non-trivial reduced connections in a midi-superspace model, we now consider spherical symmetry.

Example 4.14 (Spherical symmetry)
Outside a symmetry center of spherically symmetric spaces, we have $S = \mathrm{SU}(2)$ with $F \cong \mathrm{U}(1)$, which we realize as $F = \exp\langle\tau_3\rangle$. An invariant connection 1-form can be gauged to be $\mathbf{A} = \tilde{\mathbf{A}} + \mathbf{A}_{S/F}$ with a reduced connection 1-form $\tilde{\mathbf{A}}$ on Σ/S and

$$\mathbf{A}_{S/F} = (\Lambda(\tau_2)\sin\vartheta + \Lambda(\tau_3)\cos\vartheta)\mathrm{d}\varphi + \Lambda(\tau_1)\mathrm{d}\vartheta \tag{4.55}$$

(see Exercise 4.8). Here, (ϑ, φ) are (local) coordinates on $S/F \cong S^2$ and as usual we use the basis elements τ_i of $\mathcal{L}S$. The $\mathcal{L}F$-component $\Lambda(\tau_3)$ is given by $\mathrm{d}\lambda$, whereas $\Lambda(\tau_{1,2})$ are the scalar field components $\tilde{\phi} = \Lambda|_{\mathcal{L}F_\perp}$.

The fields Λ are not arbitrary. As in the previous example, we need to know all homomorphisms in $\mathrm{Hom}(\mathrm{U}(1), T(\mathrm{SU}(2)))$ up to conjugation, given by (4.54) with $k \in \mathbb{N}_0$. We use the homomorphisms $\lambda_k\colon \exp t\tau_3 \mapsto \exp kt\tau_3$ out of each conjugacy class. This leads to a reduced structure group $Z_G(\lambda_k(F)) = \exp\langle\tau_3\rangle \cong \mathrm{U}(1)$ for $k \neq 0$ and $Z_G(\lambda_0(F)) = \mathrm{SU}(2)$ for $k = 0$. The map $\Lambda|_{\mathcal{L}F}$ is given by $\mathrm{d}\lambda_k\colon \langle\tau_3\rangle \to \mathcal{L}G, \tau_3 \mapsto k\tau_3$, and the remaining components of Λ are subject to Eq. (4.48), or $\Lambda \circ \mathrm{ad}_{\tau_3} = \mathrm{ad}_{\mathrm{d}\lambda(\tau_3)} \circ \Lambda$. Specifically,

$$\Lambda(a_0\tau_2 - b_0\tau_1) = k(a_0[\tau_3, \Lambda(\tau_1)] + b_0[\tau_3, \Lambda(\tau_2)])$$

where $a_0\tau_1 + b_0\tau_2$, $a_0, b_0 \in \mathbb{R}$ is an element of $\mathcal{L}F_\perp$. Since a_0 and b_0 are arbitrary, we obtain two equations: $k[\tau_3, \Lambda(\tau_1)] = \Lambda(\tau_2)$ and $k[\tau_3, \Lambda(\tau_2)] = -\Lambda(\tau_1)$. A general ansatz $\Lambda(\tau_I) = a_I\tau_1 + b_I\tau_2 + c_I\tau_3$ for $I = 1, 2$ with arbitrary parameters $a_I, b_I, c_I \in \mathbb{R}$ yields

$$k(a_1\tau_2 - b_1\tau_1) = a_2\tau_1 + b_2\tau_2 + c_2\tau_3\,, \qquad k(-a_2\tau_2 + b_2\tau_1) = a_1\tau_1 + b_1\tau_2 + c_1\tau_3$$

with non-trivial solutions only for $k = 1$: $b_2 = a_1$, $a_2 = -b_1$, $c_1 = c_2 = 0$.

Now, with a non-trivial reduced manifold Σ/S given by the radial line, all free parameters are fields $a, b, c : \Sigma/S \to \mathbb{R}$. They provide the $\mathrm{U}(1)$-connection 1-form $\mathbf{A} = c(x)\,\tau_3\,\mathrm{d}x$ of the reduced connection and two scalar field components $\Lambda|_{\langle\tau_1\rangle}\colon \Sigma/S \to \mathcal{L}\mathrm{SU}(2)$, $x \mapsto a(x)\tau_1 + b(x)\tau_2$. (The general solution for Λ shows that it is sufficient to *consider* $\Lambda|_{\langle\tau_1\rangle}$ *only.) If we introduce the complex-valued field* $w(x) := \frac{1}{2}(b(x) - ia(x))$, *we have* $\Lambda(\tau_1) = \begin{pmatrix} 0 & -\overline{w}(x) \\ w(x) & 0 \end{pmatrix}$. *Under a local U(1)-gauge transformation* $z(x) = \exp(t(x)\tau_3)$ *these fields transform as* $c \mapsto c + \mathrm{d}t/\mathrm{d}x$ *and* $w(x) \mapsto \exp(-it)w$, *which can be read off from*

$$\mathbf{A} \mapsto z^{-1}\mathbf{A}z + z^{-1}\mathrm{d}z = \mathbf{A} + \tau_3 \mathrm{d}t\,,$$

$$\Lambda(\tau_1) \mapsto z^{-1}\Lambda(\tau_1)z = \begin{pmatrix} 0 & -\exp(it)\overline{w} \\ \exp(-it)w & 0 \end{pmatrix}.$$

A general invariant connection has the form (4.45), with a dual form

$$E = E^1 \sin\vartheta\, \tau_3 \frac{\partial}{\partial x} + \tfrac{1}{2}\sin\vartheta (E^2\tau_1 + E^3\tau_2)\frac{\partial}{\partial\vartheta} + \tfrac{1}{2}(E^2\tau_2 - E^3\tau_1)\frac{\partial}{\partial\varphi} \tag{4.56}$$

for an invariant densitized triad field, with coefficients E^I *canonically conjugate to* A_I. *Under local* U(1)*-gauge transformations, the radial component* $E^x := E^1$ *is invariant, while* (E^2, E^3) *transforms as a 2-vector under* $\mathrm{SO}(2) \cong \mathrm{U}(1)$. *In addition to* E^x, $E^\varphi := \sqrt{(E^1)^2 + (E^2)^2}$ *is thus invariant. Such an invariant densitized triad produces a spatial metric corresponding to (4.37) with* E^x *and* E^φ *related to* ϕ_1 *and* ϕ_2*; see Exercise 4.10. The symplectic structure*

$$\{A_I(x), E^J(y)\} = 2\gamma G \delta_I^J \delta(x, y) \tag{4.57}$$

can be derived by inserting the invariant expressions into $(8\pi\gamma G)^{-1}\int_\Sigma \mathrm{d}^3x \dot{A}_a^i E_i^a$. *Since* E^2 *and* E^3 *are non-vanishing only for* $k = 1$, *this is the unique sector providing non-degenerate triads. The topological charge which initially arises for a general spherically symmetric gauge theory (for instance, of Yang–Mills monopoles) is fixed by the requirement of a non-degenerate metric.*

After our analysis of how to classify and construct symmetric models, we now return to aspects of their dynamics.

4.2.2 Symmetric criticality

In our discussion of Bianchi class A models, we have used the Hamiltonian constraint specialized to homogeneous variables, then applied the reduced constraint to generate equations of motion. Similarly, a general reduced model is obtained by inserting the previously classified forms of invariant fields in the constraints and equations of motion. Computing equations of motion directly from the reduced constraints means that one is taking variations of a restricted expression, rather than of the full action: homogeneous configurations are only a small subspace of the infinite-dimensional space of all metrics. Since stationary points of a restricted functional supported only on a subset of all possible configurations may not agree with those of the functional on the full space, it is not guaranteed that varying a restricted action produces the correct equations of motion for the restricted configurations.

For class A models, it turns out, as shown by MacCallum and Taub (1972), that one does obtain the correct equations of motion if one first specializes the action or Hamiltonian to homogeneous variables. But there are cases where this commutation property of reduction and variation does not hold true, notably class B models.

The property that varying an action specialized to symmetric fields of some kind gives the same result as specializing the full equations of motion to this symmetry is called *symmetric criticality*. In general, there is no guarantee that this property is satisfied for a reduced model. One of the main obstructions is the presence of boundary terms. For general fields, we have treated boundaries in Chapter 3.3.2 by keeping the metric at the boundary fixed, $\delta g_{ab}|_{\partial\Sigma} = 0$. However, if the fields obey a symmetry of some form, whose orbits along which fields must remain constant intersect the boundary, variations of the boundary values are no longer independent of variations in the interior. Then, boundary terms cannot always be canceled by a surface term and differences between reduced equations of motion and equations of motion from a reduced action or Hamiltonian may arise.

Looking back at the derivation of canonical equations of motion for general relativity in ADM variables, boundary terms arose

- in the variation by δN^a: varying H_2 in (3.76), produces a boundary term from
$$\int \mathrm{d}^3x\sqrt{\det h}D_a\frac{p^{ab}\delta N_b}{\sqrt{\det h}} = \frac{1}{16\pi G}\int \mathrm{d}^3x\sqrt{\det h}D_a\left((K^{ab}-K^c_c h^{ab})\delta N_b\right)\ ;$$
- in the variation by δh_{ab}: varying H_2 produces a boundary term from
$$\int \mathrm{d}^3x\sqrt{\det h}D_f\frac{p^{cg}N^f\delta h_{cg}}{\sqrt{\det h}} = \frac{1}{16\pi G}\int \mathrm{d}^3x\sqrt{\det h}D_f\left((K^{cg}-K^a_a h^{cg})N^f\delta h_{cg}\right)$$
from one term in δG^a_{bc} (while the other terms cancel), and varying H_3 in (3.77) produces a boundary term from
$$\int \mathrm{d}^3x\sqrt{\det h}D^c\left(Nh^{de}D_c\delta h_{de} - Nh^e_c h^{df}D_d\delta h_{fe}\right)$$
$$= \int \mathrm{d}^3x\sqrt{\det h}D_a\left(N(h^{ac}h^{de}-h^{ae}h^{cd})D_c\delta h_{de}\right)$$
and one from
$$\int \mathrm{d}^3x\sqrt{\det h}D_c\left((D^cN)h^{de}\delta h_{de} - h^{cf}(D^eN)\delta h_{fe}\right)$$
$$= \int \mathrm{d}^3x\sqrt{\det h}D_c\left((D_aN)(h^{ac}h^{de}-h^{ae}h^{cd})\delta h_{de}\right)$$
due to integration by parts twice.

These terms must now specifically be checked in homogeneous models, where we cannot rely on vanishing boundary variations of the metric to make them automatically zero. However, they may vanish from the homogeneity assumption alone. This can easily be seen to be the case for the last boundary term listed because $D_aN = 0$ in any homogeneous model. To assess the other boundary terms, we make use of a reformulation of the total divergences $D_a v^a$ for vector fields as they appear in the boundary terms for different versions of v^a. In a homogeneous space, we expand v^a in the basis X_I^a of invariant vector fields, $v^a = v^I X_I^a$ with constant v^I; thus $D_a v^a = v^I D_a X_I^a$. With

$$D_a X_I^a = \frac{1}{2} h^{ab} \mathcal{L}_{X_I} h_{ab} = -\frac{1}{2} h_{ab} \mathcal{L}_{X_I} h^{ab} = -\frac{1}{2} h_{JK} \omega_a^J \omega_b^K \mathcal{L}_{X_I} (X_L^a X_M^b) h^{LM}$$
$$= -h_{JK} h^{LM} \omega_a^J \omega_b^K X_L^a \mathcal{L}_{X_I} X_M^b = -\omega_b^K [X_I, X_K]^b = -C_{IK}^K = -2a_I$$

we have

$$D_a v^a = -2v^I a_I \,. \tag{4.58}$$

The boundary terms are divergences with specific expressions for v^a. From the δN^a-variation, the boundary term becomes $a_I (K^{IJ} - K^K{}_K h^{IJ}) \delta N_J$. From the h_{ab} variation we obtain $a_I (K^{JK} - K^L{}_L h^{JK}) \delta h_{JK} N^I$ from H_2 and $a_I \omega_a^I D_c \left((h^{ac} h^{de} - h^{ae} h^{cd}) \delta h_{de} \right)$ from the first term arising from H_3. They all vanish in class A models, but not in class B models in which $a_I \neq 0$. While class A models are thus safe to use as we did — they satisfy the property of symmetric criticality — for class B models there is no straightforward reduced Hamiltonian formulation. Simply reducing the constraints of general relativity does not produce the correct reduced equations of motion in this case. (It may be possible to construct other Hamiltonian formulations not directly obtained by reducing the full action, which then do produce the correct equations of motion. But there is no systematic procedure to do so.)

There are various general results for conditions under which symmetric criticality is realized, described by Torre (1999) and Anderson et al. (2000). The main example is the compactness of the symmetry group, a criterion which applies to Bianchi IX and also to spherical symmetry. From the discussion of boundary terms, this conclusion is understandable because symmetry orbits have to extend to the boundary to have a chance of violating symmetric criticality, and then cannot be compact. Requiring staticity is an example that does not necessarily result in a correct reduced formulation: time translations correspond to non-compact symmetry orbits.

4.3 Spherical symmetry

Spherically symmetric models are the simplest inhomogeneous ones and thus provide interesting arenas in which to apply methods of general relativity. In vacuum, a unique family of solutions is given by the Schwarzschild space-time for different mass parameters. A larger class of models is provided by 2-dimensional dilaton gravity, a generalization of the spherical reduction of general relativity. Here, elegant mathematical reformulations shed light on the constraint algebra, such as the occurrence of structure functions. In 2-dimensional models, the constraint algebra can be interpreted as originating from an underlying symmetry given not by a Lie algebra as in usual gauge theories but by a Lie algebroid.

4.3.1 Schwarzschild solution

With spherical symmetry we can further illustrate the solution procedure of Einstein's equation in canonical form. As stated in the preceding section, spherical symmetry obeys symmetric criticality; we are allowed to insert the symmetric forms of variables, such as (4.45) and (4.56) in a connection formulation, into the constraints, and then determine the equations of motion that they generate.

4.3.1.1 Static solutions

We work with a spherically symmetric spatial line element. From (4.56) we see that it can only be of the form

$$\mathrm{d}s^2 = L(x)^2\mathrm{d}x^2 + S(x)^2(\mathrm{d}\vartheta^2 + \sin^2\vartheta\,\mathrm{d}\varphi^2) \tag{4.59}$$

with $L = E^\varphi/\sqrt{|E^x|}$ and $S = E^x$; see Exercise 4.10. In addition, we use a lapse function $N(x)$ and a radial shift vector with a single component $N^x(x)$ depending only on the radial coordinate[5] (as well, certainly, as on time which is suppressed in the notation). Non-vanishing connection coefficients for the spatial metric are:

$$G^x_{xx} = \frac{L'}{L}\,, \qquad G^x_{\vartheta\vartheta} = -\frac{SS'}{L^2}\,, \qquad G^x_{\varphi\varphi} = -\sin^2\vartheta\,\frac{SS'}{L^2}$$

$$G^\vartheta_{x\vartheta} = G^\varphi_{x\varphi} = \frac{S'}{S}\,, \qquad G^\varphi_{\vartheta\varphi} = \frac{\cos\vartheta}{\sin\vartheta}\,, \qquad G^\vartheta_{\varphi\varphi} = -\sin\vartheta\cos\vartheta\,.$$

Primes here denote derivatives by x. This implies the Ricci curvature components

$$^{(3)}R_{xx} = -2\frac{S''}{S} + 2\frac{S'}{S}\frac{L'}{L} \tag{4.60}$$

$$^{(3)}R_{\vartheta\vartheta} = 1 + \frac{SS'L'}{L^3} - \frac{(SS')'}{L^2} = \frac{1}{\sin^2\vartheta}\,{}^{(3)}R_{\varphi\varphi} \tag{4.61}$$

and the Ricci scalar

$$^{(3)}R = \frac{2}{S^2} - 2\frac{(S')^2}{L^2S^2} + 4\frac{S'}{S}\frac{L'}{L^3} - 4\frac{S''}{L^2S}\,. \tag{4.62}$$

Before going through a general canonical analysis, we will first be looking for static solutions, requiring vanishing momenta (as well as conditions on the frame as discussed shortly). Moreover, we choose our radial coordinate $x = r$ such that $S(r) = r$. This is a gauge-fixing condition: for spherically symmetric variables, we have the diffeomorphism constraint $D[N^x] = \int \mathrm{d}x N^x(p_S S' - Lp_L')$ with momenta p_S and p_L canonically conjugate to S and L, respectively; see the following subsection for more details. Together with $\chi := S(x) - x$, we render the diffeomorphism contribution to the constraint algebra second class: $\{D[N^x], S - x\} = -\int \mathrm{d}x N^x S'$. The diffeomorphism constraint is not just

[5] We reserve the symbol r in the context of spherical symmetry for the areal radius, which is realized if $S(r) = r$ in the metric, a specific gauge choice used below.

gauge-fixed but also, with vanishing momenta, identically satisfied; only the Hamiltonian remains to be studied.

At this stage, we implement the second condition for staticity, namely a vanishing shift vector $N^x = 0$. With this condition and vanishing momenta, the Hamiltonian constraint reads

$$0 = {}^{(3)}R = \frac{2}{r^2} - \frac{2}{r^2 L^2} + 4\frac{L'}{rL^3}\,.$$

Introducing a new function $\Lambda := L^2$, this is a differential equation

$$\Lambda' - \frac{\Lambda}{r} + \frac{\Lambda^2}{r} = 0$$

whose solutions satisfy $\Lambda/(1-\Lambda) = cr$ with a constant c, or

$$L = \sqrt{\Lambda} = \frac{1}{1+(cr)^{-1}}\,. \tag{4.63}$$

With this result, we have arrived at the correct form for all components S and L of the spatial Schwarzschild metric, but for the full space-time metric we must still determine N.

All constraints have been solved, but additional consistency conditions arise because momenta and their derivatives are assumed to vanish in this static case. Thus, with the canonical equations of motion for momenta, we must ensure that

$$\frac{16\pi G}{\sqrt{\det h}}\dot{p}^{ab} = -N\,{}^{(3)}R^{ab} + D^a D^b N - h^{ab} D_c D^c N = 0 \tag{4.64}$$

making use of the vanishing shift vector and the Hamiltonian constraint in the form ${}^{(3)}R = 0$. Equation (4.64) is indeed an equation for the lapse function.

One might wonder why such an equation arises, given that N supposedly can be chosen freely as part of the choice of frame. In general, this is true, but in the present treatment we have already made a choice for the space-time gauge by requiring the momenta to vanish. In the phase-space language, this looks like gauge-fixing conditions provided that the flow generated by the Hamiltonian constraint does not leave the condition invariant; the surface of vanishing momenta provides a cross-section of the gauge orbits generated by $H[N]$ on the constraint surface. However, we have not been so careful as to ensure proper cross-sections of the putative gauge-fixing surface (by vanishing momenta) with gauge orbits of the Hamiltonian constraint, requiring exactly one intersection per orbit. We have also imposed too many conditions: two functions, the momenta of L and S, are required to take a prescribed value even though we are now dealing with a single constraint function. An interpretation of staticity as gauge fixing is thus incorrect; instead, we view it as an ansatz: we impose staticity and look for a frame (eventually determined by the as yet unspecified lapse function) respecting this condition. By changing N, we change the Hamiltonian vector field of $H[N]$, a feature which we can exploit in trying to turn the Hamiltonian vector field parallel to the staticity surface $p_S = 0 = p_L$ in phase space. If it is not possible to turn the vector field parallel to the surface, staticity conditions cannot be preserved by the gauge flow; those conditions would have been a bad choice. If turning the vector field parallel is possible, on the other hand, staticity does provide consistent solutions, and the lapse function N realizing the parallel vector field is the consistent static frame.

The flow of the Hamiltonian constraint along its Hamiltonian vector field is exactly what we have used to derive (4.64): this equation results from the equation of motion for momenta, whose time derivatives are required to vanish. Solving this equation provides, via N and the vanishing N^a, the frame realizing this gauge choice.

In (4.64), we have[6]

$$\begin{aligned} h^{cd} D_c D_d N &= h^{rr} D_r D_r N + 2h^{\vartheta\vartheta} D_\vartheta D_\vartheta N \\ &= h^{rr}(\partial_r - G^r_{rr})\partial_r N - 2h^{\vartheta\vartheta} G^r_{\vartheta\vartheta}\partial_r N \\ &= \frac{1}{L^2}N'' - \frac{L'}{L^3}N + 2\frac{S'}{SL^2}N' \end{aligned}$$

and thus

$$\begin{aligned} (h^{rc}h^{rd} - h^{rr}h^{cd})D_c D_d N &= -2\frac{S'}{SL^4}N' \\ (h^{\vartheta c}h^{\vartheta d} - h^{\vartheta\vartheta}h^{cd})D_c D_d N &= -\frac{N''}{S^2L^2} + \frac{L'N'}{S^2L^3} - \frac{S'N'}{S^3L^2}\,. \end{aligned}$$

From the equations of motion for momenta, we obtain conditions for $a = b = r$:

$$N' - \left(\frac{S''}{S} - \frac{L'}{L}\right)N = 0$$

and for $a = b = \vartheta$:

$$-\frac{N'}{S^4}\left(1 + \frac{SS'L'}{L^3} - \frac{(SS')'}{L^2}\right) - \frac{N''}{S^2L^2} + \frac{L'N'}{S^2L^3} - \frac{S'N'}{S^3L^2} = 0\,.$$

If the last two equations are satisfied, the Hamiltonian vector field of $H[N]$ respects both conditions $p_S = 0$ and $p_L = 0$. For $S(r) = r$, the first condition implies $N'/N = -L'/L$ which is solved by $N \propto 1/L$ in agreement, combined with (4.63), with what we know from the Schwarzschild solution. At this stage, we have reproduced the full Schwarzschild line element by a canonical description, the free constant $c = -(2GM)^{-1}$ being related to the mass.

The second condition then reads

$$N'' + \left(\frac{1}{r} - \frac{L'}{L}\right)N' + \frac{L^2}{r^2}\left(1 + \frac{rL'}{L^3} - \frac{1}{L^2}\right)N = 0$$

which is consistent with the solutions already obtained for L and N; the equation need not be solved separately. While this equation does not impose further restrictions, it illustrates

[6] Note that $D_\vartheta D_\vartheta N$ does not necessarily vanish even though N depends only on r and the derivatives look deceptively angular. The abstract index notation, which may be somewhat confusing at this place, implies that $D_\vartheta D_\vartheta N$ is the ϑ-ϑ component of the tensor $D_a D_b N$. While $D_\vartheta N = 0$ as the ϑ-component of the co-vector $D_a N$, components of the tensor $D_a D_b N$ where $b = \vartheta$ do not necessarily vanish because $D_b N$ as a co-vector is non-zero.

the non-triviality of the existence of a static space-time. The staticity assumption implies that we are solving three equations for two functions, L and N. In general relativity, static solutions turn out to exist in a spherically symmetric vacuum, but this may not be the case in alternative theories of gravity in which equations of motion are changed (for instance, due to quantum effects, as discussed in the last chapter).

4.3.1.2 Reduced phase space

We have reproduced the Schwarzschild space-time in canonical form by choosing a gauge, making explicit use of the expected staticity. Illustrative is also a solution of all the canonical equations without gauge fixing, but by solving all the constraints and factoring out their gauge flows. For spherically symmetric vacuum space-times, this can be performed explicitly, as found independently by Kuchař (1994) in metric variables and by Thiemann and Kastrup (1993); Kastrup and Thiemann (1994) in complex Ashtekar variables. Bičák and Hájíček (2003) and Horváth et al. (2006) have analyzed the canonical theory of spherically symmetric space-times in the presence of null dust.

Canonical transformation Without any gauge choices, spherically symmetric general relativity in metric variables has the phase-space variables L with momentum p_L and S with momentum p_S, all depending on the radial coordinate x but not on angles. The momenta p_L and p_S can be written in terms of extrinsic-curvature components only after making use of the equations of motion for L and S, and so we view them as independent variables at this stage. From the 3-dimensional spatial perspective, they are simply components of a symmetric contravariant and spherically symmetric tensor of density weight one,

$$\mathbf{p} = \frac{1}{8\pi L} p_L \sin\vartheta \frac{\partial}{\partial x} \otimes \frac{\partial}{\partial x} + \frac{1}{16\pi S} p_S \left(\sin\vartheta \frac{\partial}{\partial\vartheta} \otimes \frac{\partial}{\partial\vartheta} + \frac{1}{\sin\vartheta} \frac{\partial}{\partial\varphi} \otimes \frac{\partial}{\partial\varphi} \right)$$

independent of the metric or its time derivatives. Choosing the tensor $\mathbf{p}$ of this form ensures its duality with the spatial metric (4.59) as dicussed in Chapter 4.2.1.3, and correct Poisson relationships between (L, S) and (p_L, p_S): we insert the spherically symmetric forms into the full 3-dimensional term $\int \mathrm{d}^3x \dot{h}_{ab} p^{ab}$ from which we read off the symplectic term

$$\int \mathrm{d}^3x \dot{h}_{ab} p^{ab} = \int \mathrm{d}x (\dot{L} p_L + \dot{S} p_S) .$$

As promised in Example 4.11, a 1-dimensional density weight must remain in the reduced variables. The density weight on spherical orbits has been removed by integrating $\sin\vartheta$ over the 2-spheres.

Inserting the spherically symmetric line element and momentum tensor into the general expressions for the Hamiltonian and diffeomorphism constraints provides their reduced forms:

$$C_{\mathrm{grav}} = -\frac{p_S p_L}{S} + \frac{L p_L^2}{2S^2} + \frac{SS''}{L} - \frac{SS'L'}{L^2} + \frac{(S')^2}{2L} - \frac{L}{2} \tag{4.65}$$

$$C_x^{\mathrm{grav}} = p_S S' - L p_L' \tag{4.66}$$

with only the radial component of the diffeomorphism constraint as a non-trivial one. (The form of the diffeomorphism constraint shows that L, not p_L, carries the density weight in the canonical pair (L, p_L); see Exercise 4.11.

Solving the constraints and factoring out the flow — providing a complete reduction to the physical, gauge-invariant degrees of freedom — requires good knowledge of the gauge flow and of observables. In spherical symmetry, we know that the mass is an observable, and as determined by Kuchař (1994) it can be expressed in canonical variables as

$$GM = \frac{p_L^2}{2S} - \frac{S(S')^2}{2L^2} + \frac{S}{2}\,. \tag{4.67}$$

Indeed, on the constraint surface this phase-space function has vanishing Poisson brackets with both constraints and is thus a Dirac observable. For the Schwarzschild solution, the momenta vanish and $L^{-2} = 1 - 2GM/x$, $S = x$ shows that M is indeed the mass. Equation (4.67), however, is valid, irrespective of the chosen gauge.

For now, we have to view GM as a phase-space function defined by (4.67), and thus as one of the fields depending on x just as L, p_L and S in its definition do. If we use it as one of the canonical variables, the further analysis should simplify, since it is left invariant by the gauge flows. We thus perform a canonical transformation to a new pair of configuration variables: the new M and the old S. This is accomplished by transforming the momentum variables to

$$p_M := \frac{Lp_L}{SF}$$

with $F := 1 - 2GM/S$, and

$$\pi_S := p_S - \frac{Lp_L}{2S} - \frac{Lp_L}{2SF} - \frac{(Lp_L)'SS' - Lp_L(SS')'}{SL^2F}\,.$$

Although we did not change the configuration variable S, it acquires a new momentum π_S to ensure that $\{GM, \pi_S\} = 0$.

The new momentum π_S turns out to be a linear combination of the constraints:

$$\pi_S = \frac{1}{F}\left(\frac{p_L C_{\rm grav}}{S} + \frac{S' C_x^{\rm grav}}{L^2}\right)\,.$$

The same property holds true for the radial derivative

$$GM' = -\frac{S' C_{\rm grav} + p_L C_x^{\rm grav}/S}{L}\,.$$

Instead of the original constraints, we can now use π_S together with M', which, at least generically, determine the same constraint surface and gauge flows thanks to the linearity of the transformation. (Some of the coefficients vanish or diverge at special subsets of the phase space, for instance one corresponding on-shell to the Schwarzschild horizon where $F = 0$. At those places, a more careful analysis is required.) A canonical action describing the system in simpler form is

$$S = \int dt \int dx (p_M G\dot{M} + \pi_S \dot{S} - N^M G M' - N^S \pi_S) \qquad (4.68)$$

with new multipliers N^M and N^S.

With constraints identical to one phase-space coordinate and the derivative of another one, the reduced-phase-space analysis is indeed much simpler than for the complicated original constraints. On the constraint surface, we have $\pi_S = 0$ with gauge flow $\delta_\epsilon S = \epsilon$. The gauge parameter ϵ can be an arbitrary function of x, allowing us to change $S(x)$ at will. The degree of freedom (S, π_S) is completely removed by implementing the constraint.

By the second constraint, $G \int dx \epsilon M'$ in smeared form, M is restricted but only required to be spatially constant. A single real number is left as phase-space degree of freedom, out of an initial functional freedom. Its momentum P_M can be changed by $\delta_\epsilon P_M = \epsilon'$. If we ignore boundaries for now, the integral $\tau := \int dx\, P_M(x)$ is gauge invariant and forms a second observable. We have now implemented all the constraints by solving them and factoring out their flows. Starting with the initial field theory on the radial line, with two functional degrees of freedom, a single canonical pair (M, τ) is left. This pair forms the reduced phase space of physical observables.

Boundary term We have so far ignored boundaries, which in the case of interest are asymptotic. A more detailed treatment will be given later in Chapter 5.3.6, but spherical symmetry already provides a glimpse on their importance. There is no finite spatial boundary in the Schwarzschild space-time, but the behavior at spatial infinity, $x \to \infty$, is of a particular form which, for the equations used here, amounts to specifying boundary conditions. More specifically, the class of space-times considered is asymptotically flat: for large values $r \gg GM$ of the areal radius on any given black-hole solution, the metric approaches the Minkowski form. Moreover, the asymptotic fall-off of metric components as functions of r is of a particular form in the Schwarzschild solution as a characteristic example: the lapse function and the radial metric component approach the Minkowski values as $1 + O(r^{-1})$, and all other components are exactly Minkowskian. For general solutions required to be asymptotically flat (though not necessarily Schwarzschild) this fall-off behavior is retained: $N(r) \sim 1 + O(r^{-1})$ and $h_{ab} = \delta_{ab} + O(r^{-1})$. To ensure that the symplectic form can be integrated, we then require momenta to fall off at least as $p^{ab} = O(r^{-2})$.

Boundary conditions have an influence on what we consider as gauge transformations. The lapse function is the multiplier of the Hamiltonian constraint; if its boundary behavior is restricted, the gauge freedom is reduced. In the present context, canonical variables are required to approach a prescribed form for $r \to \infty$. Gauge transformations are frozen as the asymptotic boundary is approached, which shows that $\int dr\, P_M$ is indeed gauge invariant even if there is a boundary on the r-axis.

Boundary conditions also play a role for the dynamics on the reduced phase space. The theory is fully constrained: there is no bulk Hamiltonian after solving all the constraints. However, the frozen gauge freedom at the asymptotic boundary implies that boundary degrees of freedom are present, as they are realized by the non-trivial reduced phase space. They are subject to dynamics, which follows from a complete action functional. So far, we have been working with (4.68) for the constraint reduction. In the presence of boundaries, however, we must be careful especially with functional derivatives by the phase-space variable M, for instance when computing gauge transformations of P_M. An integration by parts is required, leaving a boundary term $-N^M$ in the variation of S by M. The bulk term of this variation provides us with a simple equation of motion $\dot{P}_M = (N^M)'$, but the whole variation must vanish as an Euler–Lagrange equation. With a non-vanishing boundary term, this can only be achieved by adding a surface term $\int \mathrm{d}t N^M GM$ evaluated at the boundary to the original action, just as we did with the Gibbons–Hawking term for the full action. In mathematical terms, the action is functionally differentiable only when this surface term is added to it.

Physically, the consequence is that the Hamiltonian acquires the same boundary term (with a minus sign) and no longer vanishes on the constraint surface. The reduced Hamiltonian is $-N^M GM$, and thus proportional to the mass as expected for a measure of energy. Indeed, the correct equations of motion are generated on the reduced phase space: M is a constant of motion, while $\dot{\tau} = \{\tau, -N^M GM\} = N^M$ proceeds with lapse N^M.

At this stage, we have completed the reduction, showing the crucial steps of determining a reduced phase space. When this is possible, the constrained system is under full control: all observables and their reduced dynamics are known, and one may attempt a canonical quantization of the reduced canonical system if so desired. However, the present example, which from the point of view of space-time solutions — the Schwarzschild family — is not at all over complicated, also demonstrates that such complete treatments cannot be expected to be feasible in many cases. Much has to be known about the system in detailed form, and many calculational simplifications must conspire to allow solutions in closed form. Although such solvable examples are very useful as model systems (for another one, see cylindrical gravitational waves as analyzed by Kouletsis et al. (2003)), a more practical treatment of constrained systems as they arise in general relativity is given by fixing a suitable gauge.

4.3.2 2-dimensional dilaton gravity

A large class of models is obtained by formulating gravity in two space-time dimensions. Without extra fields, this does not give rise to non-trivial dynamics because the action integral $\int \mathrm{d}^2x\sqrt{-\det g}R$, according to the Gauss–Bonnet theorem, is a topological invariant and thus, despite appearances, does not depend on the metric at all. Variational equations by g_{ab} are identically satisfied and no space-time dynamics ensues.

Coupling a scalar field to gravity in two dimensions, introducing the dilaton ϕ in the form

$$S = \frac{1}{4G}\int_\Sigma \mathrm{d}^2x\sqrt{-\det g}(\phi R - V(\phi)) \tag{4.69}$$

not only results in non-trivial dynamics but also corresponds to several cases of physical interest depending on the choice of the dilaton potential $V(\phi)$ (as reviewed by Grumiller et al. (2002)). For $\phi = S^2$, $V(\phi) = -2\sqrt{\phi}$, for instance, one obtains the action of spherically reduced gravity with S as it appears in the line element (4.59). The 2-dimensional metric and Ricci scalar then refer to the (t, r)-part of the line element, rescaled by a conformal factor S. (Reducing D-dimensional gravity to spherically symmetric form results in a potential $V(\phi) = -(D-2)(D-3)\phi^{1/(D-2)}$ in $D > 3$ space-time dimensions, with $\phi = S^{D-2}$. The conformal transformation is then done with S^{D-3}.) Finally, 2-dimensional dilaton gravity models are included in the more general class of Poisson sigma models, introduced by Ikeda and Izawa (1993); Ikeda (1994) and Schaller and Strobl (1994), which show a rich amount of interesting symmetries and mathematical structures.

4.3.2.1 Poisson sigma model

To reformulate (4.69) and introduce Poisson sigma models, we express the action in first-order form using co-dyads e_a^α and connection 1-forms $\omega_a{}^\alpha{}_\beta$. Indices $\alpha, \beta, \ldots$ refer to the internal 2-dimensional Minkowski spaces on which the 1-dimensional structure group $\mathrm{SO}(1,1) \cong \mathbb{R}$ is acting. We will mainly be using light-cone indices $\pm$, corresponding to the coordinates $x^\pm := x \pm t$. Co-dyad components with respect to this coordinate basis are $e_a^\pm = 2^{-1/2}(e_a^0 \pm e_a^1)$.

In two dimensions, the curvature scalar can be realized in a simple way. Starting with the Riemann tensor, tensorial symmetries show that it must be characterized completely by a single function X such that $R_{ab\alpha}{}^\beta = \sqrt{-\det g}\, X \varepsilon_{ab} \epsilon_\alpha{}^\beta$. Here, ε_{ab} is the tensor density of density weight -1, which we remove by multiplying with the determinant of the 2-dimensional metric.

Example 4.15 (2-dimensional curvature tensors)
For a Riemann tensor $R_{ab\alpha}{}^\beta = \sqrt{-\det g}\, X \varepsilon_{ab} \epsilon_\alpha{}^\beta$ on a 2-dimensional space-time we have the Ricci tensor

$$\begin{aligned} R_{ab} &= e_b^\alpha e_\beta^c R_{ac\alpha}{}^\beta = \sqrt{-\det g}\, X e_b^\alpha e_\beta^c \varepsilon_{ac} \epsilon_\alpha{}^\beta \\ &= \sqrt{-\det g}\, X \varepsilon_{ac} g_{bd} \varepsilon^{dc} (\det e)^{-1} = -X g_{ab} \end{aligned}$$

using the determinant identity $\epsilon^{\alpha\beta} e_\alpha^d e_\beta^c = (\det e)^{-1} \varepsilon^{dc}$ with the determinant $\det e$ of the co-dyad e_a^α. The Ricci scalar is then obtained as $R = -2X$; thus, $R_{ab} = \frac{1}{2} R g_{ab}$. In two space-time dimensions, the Einstein tensor $G_{ab} = R_{ab} - \frac{1}{2} R g_{ab}$ vanishes identically. 2-dimensional gravity becomes non-trivial only by coupling the dilaton field, or other matter fields.

Like the curvature tensor, the Ricci rotation coefficients $\omega_a{}^\alpha{}_\beta$ can also be reduced to a single 1-form ω_a such that $\omega_a{}^\alpha{}_\beta = \omega_a \epsilon^\alpha{}_\beta$. Again, this reduction follows from the tensorial symmetries, or from the fact that we have a 1-dimensional Lie algebra so(1, 1) with a single generator $\epsilon^\alpha{}_\beta$. Also owing to the 1-dimensionality of the structure group, the second structure equation (3.134) simplifies: $\mathbf{R}_\alpha{}^\beta = \mathrm{d}\boldsymbol{\omega}_\alpha{}^\beta$. Together with our expression for the

Riemann tensor in terms of $X = -\frac{1}{2}R$, this implies that $-\frac{1}{2}R\sqrt{-\det g}\varepsilon_{ab}\epsilon_\alpha{}^\beta = (\mathrm{d}\boldsymbol{\omega})_{ab}\epsilon_\alpha{}^\beta$, or $R\boldsymbol{\epsilon} = -2\mathrm{d}\boldsymbol{\omega}$ with the volume form $\boldsymbol{\epsilon}_{ab} = \sqrt{-\det g}\varepsilon_{ab}$.

We can thus write the action as

$$S = \frac{1}{4G}\int_\Sigma \mathrm{d}^2x\sqrt{-\det g}(\phi R - V(\phi)) = -\frac{1}{2G}\int_\Sigma\left(\phi\mathrm{d}\boldsymbol{\omega} + \frac{1}{2}V(\phi)\boldsymbol{\epsilon}\right) \tag{4.70}$$

where ω_a is understood as the connection 1-form compatible with the dyad, satisfying $\mathcal{D}\mathbf{e}^\alpha = \mathrm{d}\mathbf{e}^\alpha + \boldsymbol{\omega}\epsilon^\alpha{}_\beta \wedge \mathbf{e}^\beta = 0$ according to the first structure equation (3.133). We are going to use this action also in first-order form, but unlike in the 4-dimensional case the first structure equation in two dimensions is not implied by the equations of motion: by variations of the action (4.70), ω_a is restricted only by the second structure equation. In order to ensure the correct connection 1-form in a first-order form, we add the first structure equation $\mathcal{D}\mathbf{e}^\alpha = 0$ as a constraint, with Lagrange multipliers X_α:

$$S = -\frac{1}{2G}\int_\Sigma\left(\phi\mathrm{d}\boldsymbol{\omega} + \frac{1}{2}V(\phi)\boldsymbol{\epsilon} + X_\alpha\mathcal{D}\mathbf{e}^\alpha\right). \tag{4.71}$$

This is the first-order action for dilaton gravity. (If the dilaton potential is allowed to depend on the Lorentz invariant $X^\alpha X_\alpha$ in addition to ϕ, a theory with torsion results.) Integrating by parts (discarding boundary terms), we arrive at the action

$$S = -\frac{1}{2G}\int_\Sigma\left(\mathbf{e}^\alpha \wedge \mathrm{d}X_\alpha + \boldsymbol{\omega}\wedge\mathrm{d}\phi + X_\alpha\varepsilon^\alpha{}_\beta\boldsymbol{\omega}\wedge\mathbf{e}^\beta + \frac{1}{2}V(\phi)\boldsymbol{\epsilon}\right).$$

Renaming the various variables as

$$(X^i)_{i=1,2,3} := (X_\alpha, \phi) = (X_+, X_-, \phi) \tag{4.72}$$

for the scalars, and

$$(\mathbf{A}_i)_{i=1,2,3} := (\mathbf{e}^\alpha, \boldsymbol{\omega}) = (\mathbf{e}^+, \mathbf{e}^-, \boldsymbol{\omega}) \tag{4.73}$$

for the 1-forms, and introducing the matrix

$$(\mathcal{P}^{ij}) = \begin{pmatrix} 0 & -\frac{1}{2}V(\phi) & -X_- \\ \frac{1}{2}V(\phi) & 0 & X_+ \\ X_- & -X_+ & 0 \end{pmatrix} \tag{4.74}$$

the action becomes

$$S = -\frac{1}{2G}\int_\Sigma\left(\mathbf{A}_i \wedge \mathrm{d}X^i + \frac{1}{2}\mathcal{P}^{ij}\mathbf{A}_i\wedge\mathbf{A}_j\right). \tag{4.75}$$

We now have a connection formulation with 3-dimensional internal spaces, coordinatized by X^i: compared with space-time, the dimension is enlarged since the dilaton and the connection have been included as additional internal directions by the definitions (4.72), (4.73). No new symmetries on the internal space arise from this extension, but the internal space does acquire a new structure: it has become a Poisson manifold, since the matrix (4.74), interpreted as a bivector, satisfies all conditions for a Poisson tensor; see Exercise 4.14.

The action thus describes a sigma model of fields $X^i : \Sigma \to (M, \mathcal{P})$, $x^\alpha \mapsto X^i(x^\alpha)$ whose world-sheet Σ is 2-dimensional and whose target space M is a 3-dimensional Poisson manifold: a *Poisson sigma model*. Later on, we will be able to interpret the connection fields $\mathbf{A}_i$ as a related mapping of algebraic structures.

Field equations are easily obtained by variation:

$$-2G\frac{\delta S}{\delta \mathbf{A}_i} = \mathrm{d}X^i + \mathcal{P}^{ij}(X)\mathbf{A}_j = 0 \tag{4.76}$$

$$-2G\frac{\delta S}{\delta X^i} = \mathrm{d}\mathbf{A}_i + \frac{1}{2}\left(\partial_i \mathcal{P}^{kl}(X)\right)\mathbf{A}_k \wedge \mathbf{A}_l = 0\,. \tag{4.77}$$

Moreover, the action is invariant under the local transformations

$$X^i \mapsto X^i + \epsilon_j \mathcal{P}^{ji}(X) \quad , \quad \mathbf{A}_i \mapsto \mathbf{A}_i + \mathrm{d}\epsilon_i + \partial_i \mathcal{P}^{kl}(X)\mathbf{A}_k \epsilon_l \tag{4.78}$$

for arbitrary functions ϵ_i. (This symmetry relies on $\mathcal{P}^{ij}$ being a Poisson tensor, satisfying the Jacobi identity. The Poisson property is thus crucial for Poisson sigma models.)

Poisson sigma models are more general than dilaton gravity. Other examples are:

Example 4.16 (Gauge theories as Poisson sigma models)

1. *BF-theory: If $\mathcal{P}^{ij}$ is the Poisson tensor defined on the dual of a Lie algebra $\mathcal{L}G$ by the Kirillov–Kostant structure, $\mathcal{P}^{ij} = C^{ij}{}_k X^k$ with the structure constants $C^{ij}{}_k$, the action of the resulting Poisson sigma model can be written compactly as*

$$S(X, \mathbf{A}) = \int_\Sigma X \cdot F(\mathbf{A})$$

with the $\mathcal{L}G$-valued curvature 2-form $F(\mathbf{A}) = \mathrm{d}\mathbf{A} + \frac{1}{2}C^{ij}{}_k \mathbf{A}_i \wedge \mathbf{A}_j T^k = \mathrm{d}\mathbf{A} + [\mathbf{A}, \mathbf{A}]$. (The form of the Lagrangian density, the field X often called B, gives rise to the name BF-theory.) While $F(\mathbf{A})$ takes values in $\mathcal{L}G$, X takes values in the dual $(\mathcal{L}G)^*$. Applying a dual element to one in $\mathcal{L}G$ is indicated by the '$\cdot$'. The field equations are $\mathcal{D}_\mathbf{A} X = \mathrm{d}X + [\mathbf{A}, X] = 0$ and $F(\mathbf{A}) = 0$. Symmetries (4.78) of the action integrate to the standard gauge transformations $X \mapsto g^{-1}Xg$ and $\mathbf{A} \mapsto g^{-1}\mathbf{A}g + g^{-1}\mathrm{d}g$ with $g = \exp(\epsilon_i T^i) \in G$ in the Lie group.
2. *Yang–Mills theory is obtained in a similar way after adding $C(X)\mathbf{w}$ to the action with C the quadratic Casimir function of the Lie algebra and $\mathbf{w}$ a volume form on the world-sheet. In contrast to BF-theory, the field equations of Yang–Mills theory in two dimensions thus partially depend on a background geometry via the volume form.*

In general, Poisson sigma models allow arbitrary dimensions for the target space involved. For dilaton gravity, the target space is 3-dimensional, but coupling BF-theory or Yang–Mills theory to gravity increases the dimension. From now on, we will work with general target spaces required only to carry a Poisson structure.

4.3.2.2 Solutions

Solutions of a Poisson sigma model in general terms can be characterized entirely by the target manifold's Poisson geometry. Most importantly, the structure of symplectic leaves which foliate the Poisson manifold is important. (See the Appendix or Chapter 3.1.2.2.)

For a symplectic manifold, there is only one symplectic leaf, the whole manifold. For a degenerate Poisson tensor, however, symplectic leaves are of lower dimensions than the full manifold and defined in such a way that they carry a symplectic structure compatible with the Poisson tensor. One can construct the leaves using the subbundle $\mathcal{P}^\sharp(T^*M) \subset TM$, which defines an integrable distribution owing to the Jacobi identity. Integrating the distribution results in a foliation by the symplectic leaves.

Since tangent spaces to the leaves are in the image of $\mathcal{P}^\sharp$, they carry a non-degenerate symplectic structure. The co-normal space to the leaves is $\ker\mathcal{P}^\sharp$, and can locally be given in terms of Casimir functions C^I such that $\mathcal{P}^\sharp(\mathrm{d}C^I) = \mathcal{P}^{ij}\partial_i C^I \partial_j = 0$ as a vector field. Still locally, one can always introduce Casimir–Darboux coordinates[7] (x^α, C^I) such that the Poisson tensor takes the standard form (A.13) given in the Appendix, introducing the leaf-symplectic structure $\Omega_{\alpha\beta}$.

Writing the field equations of a Poisson sigma model in Casimir–Darboux coordinates for the target space, they reduce to

$$\mathrm{d}X^I = 0\,, \qquad \mathrm{d}X^\alpha = \mathcal{P}^{\alpha\beta}\mathbf{A}_\beta\,, \qquad \mathrm{d}\mathbf{A}_I = 0\,.$$

The equation for $\mathrm{d}A_\alpha$ is solved automatically by inverting $\mathrm{d}X^\alpha = \mathcal{P}^{\alpha\beta}\mathbf{A}_\beta$ to $\mathbf{A}_\alpha = \Omega_{\alpha\beta}\mathrm{d}X^\beta$. Gauge transformations (4.78) take the form

$$\delta_\epsilon X^I = 0\,, \qquad \delta_\epsilon X^\alpha = \epsilon_\beta \mathcal{P}^{\beta\alpha}\,, \qquad \delta_\epsilon \mathbf{A}_I = \mathrm{d}\epsilon_I\,, \qquad \delta_\epsilon \mathbf{A}_\alpha = \mathrm{d}\epsilon_\alpha\,.$$

We first solve the equations for X^i: its components must satisfy $X^I = \mathrm{const}^I$, which means that the image of Σ under X must entirely lie in one symplectic leaf of $(M, \mathcal{P})$ given by the Casimir values $C^I = X^I$. The other components X^α are undetermined and in fact can be changed arbitrarily by choosing a gauge: on the leaf, $\mathcal{P}^{\alpha\beta}$ is invertible such that $\delta_\epsilon X^\alpha = \epsilon_\beta \mathcal{P}^{\beta\alpha}$ includes all vector fields tangent to the leaf. The image of Σ under X^i can thus be deformed within the leaf without restrictions; gauge invariant is only the homotopy class of the mapping.

The connection components $\mathbf{A}_\alpha$ are determined from X^α while the gauge transformation for $\mathbf{A}_I$ tells us that gauge-invariant solutions are given by elements $\boldsymbol{\alpha}_I$ of the first cohomology group of Σ: the space of closed 1-forms on Σ, $\mathrm{d}\mathbf{A}_I = 0$, modulo exact ones, identifying $\mathbf{A}_I$ with $\mathbf{A}_I + \mathrm{d}\epsilon_I$. If the leaf allows a compatible presymplectic form $\tilde{\Omega}_{ij}$, one can, following Bojowald and Strobl (2003a), write a complete solution for $\mathbf{A}_i$ as

$$\mathbf{A}_i = -X^*(\tilde{\Omega}_{ij}\mathrm{d}X^j) + \boldsymbol{\alpha}_I X^*(\partial_i C^I)\,.$$

4.3.2.3 Canonical formulation

The action (4.75) is already in first-order form and can easily be decomposed in a space-time splitting:

$$S = -\frac{1}{2G}\int \mathrm{d}t \int \mathrm{d}x \left(A_i \dot{X}^i - \Lambda_i((X^i)' + \mathcal{P}^{ij}A_j)\right) \tag{4.79}$$

[7] We are recycling the letters $\alpha, \beta, \ldots$ for indices; internal 2-dimensional Minkowski indices will no longer be used from now on.

with

$$A_i := \mathbf{A}_{xi} \quad \text{and} \quad \Lambda_i := \mathbf{A}_{ti}\,. \tag{4.80}$$

We immediately read off the canonical variables with their field-space Poisson brackets

$$\{A_j(x), X^i(y)\} = 2G\delta^i_j\delta(x, y)$$

(to be distinguished from the Poisson structure given by $\mathcal{P}^{ij}$ on the finite-dimensional target space). Since Λ_i does not appear with time derivatives, its momenta vanish as primary constraints, implying the secondary constraint $\tilde{G}[\Lambda_i] = \int \mathrm{d}x\, \Lambda_i \tilde{G}^i \approx 0$ with

$$\tilde{G}^i := -\frac{1}{2G}((X^i)' + \mathcal{P}^{ij}A_j)\,.$$

With infinitesimal parameters ϵ_j, the constraints $\tilde{G}[\epsilon_j]$ generate gauge transformations

$$\delta_\epsilon X^i = \epsilon_j \mathcal{P}^{ji}(X) \tag{4.81}$$

$$\delta_\epsilon A_i = \mathrm{d}\epsilon_i + \frac{1}{2}(\partial_i \mathcal{P}^{kl}(X))A_k\epsilon_l \tag{4.82}$$

in agreement with (4.78).

The constraints can easily be seen to form the first-class algebra

$$\{\tilde{G}[\Lambda_i], \tilde{G}[K_j]\} = -\frac{2}{G}\tilde{G}[\Lambda_i K_l \partial_k \mathcal{P}^{il}]\,. \tag{4.83}$$

For the Poisson sigma model corresponding to dilaton gravity, the constraints include generators of spatial diffeomorphisms, seen by noting that the combination $D := -A_i\tilde{G}^i = A_i(X^i)'$ has the correct form for a diffeomorphism constraint in one spatial dimension; and indeed, the transformations generated by D are Lie derivatives of the fields. The constraint $\tilde{G}^3$ generates local so(1, 1)-gauge transformations, and the remaining independent combination of $\tilde{G}^i$ amounts to a Hamiltonian constraint. In Poisson sigma models, all constraints, including the dynamical ones, can be described by a single algebra in terms of the Poisson tensor. We will exploit this convenient feature in the next subsection.

Example 4.17 (Linear Poisson tensor)
For a linear Poisson tensor $\mathcal{P}^{ij} = C^{ij}_k X^k$, the constraint algebra (4.83) has structure constants $\partial_k \mathcal{P}^{ij} = C^{ij}_k$, rather than structure functions that would depend on X^i. The constraints are equivalent to the usual Gauss constraint of a gauge theory whose structure group has the same structure constants. Indeed, as already seen in the context of BF-theory in Example 4.16, the gauge transformations (4.78) in this case are of the well-known form

$$\delta_\epsilon X^i = C^{ji}_k \epsilon_j X^k = [\epsilon, X]^i \tag{4.84}$$

$$\delta_\epsilon A_i = \mathrm{d}\epsilon_i + C^{kl}_i A_k \epsilon_l = \mathrm{d}\epsilon_i - [\epsilon, A]_i \tag{4.85}$$

with the Lie bracket $[X, Y]_k = C^{ij}_k X_i X_j$.

For non-linear Poisson tensors as realized, for instance, by spherically symmetric gravity, on the other hand, the system has structure functions. With this property, non-linear Poisson sigma models provide interesting systems exhibiting one of the key aspects of full general relativity. Even in the case of structure functions, the gauge structure of Poisson sigma models can still be understood algebraically. However, they no longer come from a Lie algebra but from the more general notion of a Lie algebroid.

4.3.2.4 Lie algebroids

As already seen, for a linear Poisson tensor the fields X^i take values in the dual of a Lie algebra, and $\mathbf{A}_i$ maps vector fields on the world-sheet to co-vectors on the dual Lie algebra or, using double duality, to elements of the Lie algebra itself. In the non-linear case, a Poisson manifold $(M, \mathcal{P})$ does not have an interpretation as the dual of a Lie algebra, but it forms a *Lie algebroid* as defined in the Appendix: a vector bundle $E = T^*M$ over the Poisson manifold M with an *anchor map* $\rho\colon T^*M \to TM, \alpha_i \mapsto \mathcal{P}^{ij}\alpha_j$ and a *Lie bracket* $[\mathrm{d}X^i, \mathrm{d}X^j] = \partial_k\mathcal{P}^{ij}\mathrm{d}X^k$ on coordinate differentials on M, extended to all sections of T^*M by linearity and the Leibniz rule. Compatibility conditions between the anchor map and the Lie bracket are satisfied, making the co-tangent bundle of a Poisson manifold a Lie algebroid. A Lie algebra is a Lie algebroid over a single point, and so Lie algebroids naturally generalize Lie algebras.

The tangent space of a manifold is also a Lie algebroid, with the anchor being the identity map and the usual Lie bracket of vector fields. The fields of a Poisson sigma model can be interpreted as maps between these two Lie algebroids: the first one given by the tangent bundle of the world-sheet Σ with the usual Lie bracket of vector fields, the second one as the co-tangent bundle of the target manifold M with its Lie bracket defined via the Poisson tensor. Any section of $E_1 := T\Sigma$, i.e. a vector field v^a, is mapped to a section of $E_2 := T^*M$ with co-vector $A_{ia}v^a(x)$ at $X(x)$. When the fields satisfy the equations of motion of a Poisson sigma model, these maps are Lie algebroid morphisms mapping the structures of the Lie algebroids, anchor and Lie bracket, into each other. The specific laws to be satisfied by a Lie algebroid morphism are collected in the Appendix.

This statement can be proved by direct calculations. We start by showing the first property of a Lie algebroid morphism $\phi\colon E_1 \to E_2$, namely $\rho_2 \circ \phi = (\phi_0)_* \circ \rho_1$. Here, $\rho_{T\Sigma} = \mathrm{id}$, $\rho_{T^*M} = \mathcal{P}^\sharp$, $\phi_0 = X$ and $\phi(v) = A_{ia}v^a\mathrm{d}X^i \in T^*_{X(x)}M$ for any $v \in T_x\Sigma$. Thus,

$$(\rho_{T^*M} \circ \phi)(v) = \mathcal{P}^{ji}A_{ja}v^a\partial_i$$

and

$$((\phi_0)_* \circ \rho_{T\Sigma})(v) = X_*v = \mathrm{d}X(v) = \partial_a X^i v^a \partial_i$$

by the chain rule. Both expressions are equal if and only if $\mathrm{d}X^i = \mathcal{P}^{ji}\mathbf{A}_j$, which is the first field equation (4.76).

The second property of a Lie algebroid morphism then ensures that, in addition to the anchors, the algebraic operations are also respected: for sections $s_{1/2}$ and $s'_{1/2}$ of $E_{1/2}$, the conditions $\phi \circ s_1 = s_2 \circ \phi_0$ and $\phi \circ s'_1 = s'_2 \circ \phi_0$ must ensure $\phi \circ [s_1, s'_1]_1 = [s_2, s'_2]_2 \circ \phi_0$.

In our case, the assumption means that for all vector fields v^a and w^a on Σ, there are sections α_i and β_i of T^*M such that $A_{ia}v^a = X^*\alpha_i$ and $A_{ia}w^a = X^*\beta_i$ as sections of $X^*(T^*M)$, i.e. as fields on Σ taking values in T^*M. For $\phi \circ [s_1, s_1']_1$, we compute $A_{ia}[v, w]^a$, which can easily be done using Cartan's formula (see Exercise A.3 in the Appendix):

$$A_{ia}[v, w]^a = (\mathrm{d}\mathbf{A}_i)_{ab} w^a v^b + v^a \partial_a(A_{ib}w^b) - w^a \partial_a(A_{ib}v^b)\,.$$

For the terms on the right-hand side, we have

$$\begin{aligned}(\mathrm{d}\mathbf{A}_i)_{ab} w^a v^b &= -\frac{1}{2}\partial_i \mathcal{P}^{kl} A_{ka} A_{lb} w^a v^b = -X^*(\partial_i \mathcal{P}^{kl} \beta_k \alpha_l)\\ &= X^*\left(([\mathrm{d}X^k, \mathrm{d}X^l]_{T^*M})_i \alpha_k \beta_l\right)\end{aligned}$$

using the second field equation (4.77) and the definitions. Then,

$$v^a \partial_a(A_{ib}w^b) = X^*\left((X_* v)\beta_i\right) = X^*\left((\mathcal{P}^\sharp \alpha)\beta_i\right)$$

where we have used

$$\mathcal{P}^\sharp(X^*\alpha) = \mathcal{P}^{ij}(X^*\alpha_i)\partial_j = \mathcal{P}^{ij} A_{ia} v^a \partial_j = (\mathrm{d}X^j)_a v^a \partial_j = \partial_a X^j v^a \partial_j = X_* v$$

with the first field equation. Fully analogously,

$$w^a \partial_a(A_{ib}v^b) = X^*\left((\mathcal{P}^\sharp \beta)\alpha_i\right)\,.$$

Combining these equations gives

$$A_{ia}[v, w]^a = X^*\left(([\mathrm{d}X^k, \mathrm{d}X^l]_{T^*M})_i \alpha_k \beta_l + \rho(\alpha)\beta_i - \rho(\beta)\alpha_i\right) = X^*([\alpha, \beta]_i)$$

according to the Leibniz rule. Since A_{ia} represents ϕ, and X^i provides ϕ_0, preservation of the brackets follows from the field equations, and vice versa.

4.3.2.5 Symmetries

Solutions $(X^i, \mathbf{A}_i)$ of the field equations for a Poisson sigma model are in one-to-one correspondence with Lie algebroid morphisms from the tangent space of the world-sheet to the tangent space of the target Poisson manifold. Symmetries, with structure functions in the non-linear case, act on the fields and should be interpreted as relationships between Lie algebroid morphisms. In the linear case, symmetries can be realized just on the target space, by the usual pointwise gauge transformations from the Lie algebra. In the non-linear case, world-sheet and target manifold can no longer be split.

We will denote the anchor map components as ρ_I^i, as they arise from applying the anchor map to basis sections b_I of E: $\rho(b_I) =: \rho_I^i \partial_i$. For the structure functions, we write C^I_{JK}, obtained from $[b_J, b_K] = C^I_{JK} b_I$. All these components are determined by the Poisson tensor in the case of a Poisson sigma model, but it is useful to denote them by their algebraic meaning. At the same time, a generalization to Lie algebroid morphisms not related to Poisson manifolds opens up in this way, as introduced by Strobl (2004). Our symmetries then read $\delta_\epsilon X^i = \rho_I^i \epsilon^I$ and $\delta_\epsilon \mathbf{A}^I = \mathrm{d}\epsilon^I + C^I_{JK}\mathbf{A}^J \epsilon^K$ with a section ϵ^I of E_2. The pair $(X^i, \mathbf{A}^I)$ provides a Lie algebroid morphism; our field space is thus $\mathcal{M} = \{\phi\colon E_1 \to E_2\}$, while X^i alone provides maps in $\mathcal{M}_0 = \{\phi_0 \colon \Sigma \to M\}$. From $\delta_\epsilon X^i$ we obtain a vector field on $\mathcal{M}_0$, but including $\delta_\epsilon \mathbf{A}^I$ does not automatically result in a vector field on $\mathcal{M}$. The definition of $\delta_\epsilon \mathbf{A}^I$ is not frame-covariant: if we change our basis for sections of E, such that the components of the gauge generator ϵ^I and of $\mathbf{A}^I$ itself transform according to $\epsilon^I = B^I_J(X)\tilde{\epsilon}^J$, the symmetry transformation $\mathrm{d}\epsilon^I + C^I_{JK}\mathbf{A}^J\epsilon^K$ of $\mathbf{A}^I$ does not transform as $\mathbf{A}^I$ does. By the Leibniz rule, we should have

$$\delta_\epsilon(B^I_J\tilde{\mathbf{A}}^J) = B^I_J\delta_\epsilon\tilde{\mathbf{A}}_J + \tilde{\mathbf{A}}^J(\partial_i B^I_J)\delta_\epsilon X^i$$

but what we find for the symmetry transformation is

$$\begin{aligned}\mathrm{d}\epsilon^I + C^I_{JK}\mathbf{A}^J\epsilon^K &= B^I_J(\mathrm{d}\tilde{\epsilon}^I + \tilde{C}^I_{JK}\tilde{\mathbf{A}}^J\tilde{\epsilon}^K)\\ &\qquad + (\partial_i B^I_J)\mathrm{d}X^i\tilde{\epsilon}^J + (\partial_i B^I_J)\tilde{\rho}^i_K\tilde{\mathbf{A}}^J\tilde{\epsilon}^K - (\partial_i B^I_J)\tilde{\rho}^i_K\tilde{\mathbf{A}}^K\tilde{\epsilon}^J\end{aligned}$$

with $B^M_K\rho^i_M =: \tilde{\rho}^i_K$. This equation is obtained using

$$C^I_{LM}B^L_J B^M_K = \tilde{C}^N_{JK}B^I_N + \tilde{\rho}^i_K\partial_i B^I_J - \tilde{\rho}^i_J\partial_i B^I_K$$

which in turn follows from the definitions $[b_J, b_K] = C^I_{JK}b_I$ and $[B^L_J b_L,\ B^M_K b_M] = \tilde{C}^N_{JK}B^I_N b_I$ and the Leibniz rule $[s, fs'] = f[s, s'] + (\rho(s)f)s'$ of a Lie algebroid. The difference of all terms in the frame-transformed symmetries is $(\partial_i B^I_J)(\mathrm{d}X^i - \rho^i_I\mathbf{A}^I) =: (\partial_i B^I_J)\mathbf{F}^i\tilde{\epsilon}^J$ which is not identically zero, but vanishes on-shell. We thus expect that the field equations must be used to interpret the symmetries as relationships between Lie algebroid morphisms. We will sketch this proof for on-shell symmetries following Bojowald et al. (2005b), who also have provided a more complicated extension to off-shell symmetries.

Definition 4.1 *Two Lie algebroid morphisms $\phi, \phi' \colon E_1 \to E_2$ are* homotopic *if and only if there is a morphism $\bar{\phi}$ from the direct sum $\bar{E} := E_1 \oplus TI$ over $M_1 \times I$ with the interval $I = [0, 1]$ to the Lie algebroid E_2 such that $\bar{\phi}$ restricted to $M_1 \times \{0\}$ equals ϕ_1, and $\bar{\phi}$ restricted to $M_1 \times \{1\}$ equals ϕ'.*

Lie algebroid morphisms resulting from solutions to a Poisson sigma model over the world-sheet Σ can naturally be extended to the tangent bundle over $\Sigma \times I$: With a coordinate $t \in I$, fields $\bar{X}^i(x, t)$ and $\bar{\mathbf{A}}^I = \mathbf{A}^I + A^I_t\mathrm{d}t$ provide a general extension $\bar{\phi}$. We will denote the differential operator on $\Sigma \times I$ as $\bar{\mathrm{d}} = \mathrm{d} + \mathrm{d}t \wedge \partial_t$. The structure functions C^I_{JK} and anchor map components ρ^i_I remain unchanged and t-independent, since they refer only to E_2.

For $\bar{\phi}$ to be a morphism, the field equations of a Poisson sigma model must be satisfied for the components $\bar{X}^i$ and $\bar{\mathbf{A}}^I$ defining the extended morphism. Thus, we must have

$$0 = \bar{\mathbf{F}}^i := \bar{\mathrm{d}}X^i - \rho^i_I\bar{\mathbf{A}}^I = \mathbf{F}^i(t) + \mathrm{d}t \wedge (\partial_t X^i - \rho^i_I A^I_t)$$

using the non-extended $\mathbf{F}^i$ at each t, and

$$0 = \bar{\mathbf{F}}^I := \bar{\mathrm{d}}\bar{\mathbf{A}}^I + \frac{1}{2}C^I_{JK}\bar{\mathbf{A}}^J \wedge \bar{\mathbf{A}}^K = \mathbf{F}^I(t) + \mathrm{d}t \wedge (\partial_t\mathbf{A}^I - \mathrm{d}A^I_t + C^I_{JK}A^J_t\mathbf{A}^K)\,.$$

Identifying $\epsilon^I := A_t^I$ shows that the field equations are satisfied for $\bar{\phi}$ if and only if they are satisfied for all fixed t and $\partial_t X^i = \delta_\epsilon X^i$, $\partial_t \mathbf{A}^I = \delta_\epsilon \mathbf{A}^I$. Thus, we have a homotopy of Lie algebroid morphisms, represented by two different solutions of a Poisson sigma model, if and only if the solutions are related by the symmetries of a Poisson sigma model.

4.3.2.6 The gauge groupoid of canonical gravity

Poisson sigma models allow one to formulate the gauge transformations (with structure functions) underlying spherically symmetric gravity as the action of a Lie algebroid. As shown by Cattaneo and Felder (2000), the class of Lie algebroids encountered here can be integrated to Lie groupoids. Also, the full gauge algebra (3.88), (3.89) and (3.90) can be interpreted as a Lie algebroid, derived from a Lie groupoid, but the realization is much less direct than in the case of spherical symmetry.

Blohmann et al. (2010) show that pairs (N, N^a) of a function and a vector field on a spatial manifold Σ form a Lie algebroid over the base manifold $\mathcal{M}$ of metrics h_{ab} on Σ with the bracket

$$[(N, N^a), (M, M^a)] = (N^a \partial_a M - M^a \partial_a N, [N, M]^a - h^{ab}(N \partial_b M - M \partial_b N))$$

and anchor map

$$\rho(N, N^a)_{h_{ab}} = \mathcal{L}_{t_{(N,N^a)}} h_{ab}$$

with a vector field $t^a_{(N,N^b)} = N n^a + N^a$ for an embedding of Σ in space-time such that it acquires the normal n^a and the space-time metric becomes Gaussian. The construction thus requires a choice of gauge, amounting to a local trivialization of the Lie algebroid constructed. Globally, however, the Lie algebroid exists without requiring gauge choices.

The Lie algebroid is the infinitesimal version of a Lie groupoid which can be described in simple terms. Consider as objects all space-time evolutions $M_{1\to 2}$ from one spatial slice Σ_1 to another one Σ_2 according to Einstein's equation. We identify $s(M_{1\to 2}) := \Sigma_1$ as the source and $t(M_{1\to 2}) := \Sigma_2$ as the target of M. These objects can be composed by adjoining the space-time regions, provided that the target of the first one equals the source of the second one, in this way defining a groupoid. As Blohmann et al. (2010) demonstrate, this is the Lie groupoid integrating the Lie algebroid underlying the constraint algebra.

4.4 Linearized gravity

When the gravitational field is not too strong, Einstein's equation can be linearized. This is often used to discuss gravitational waves, cosmological structure formation or physics near the Newtonian limit. Also, the stability of properties found in symmetric models can be analyzed by introducing perturbations around them. In those cases, one starts with a background metric ${}^0g_{ab}$ around which the metric is defined as a perturbation $g_{ab} = {}^0g_{ab} + \delta g_{ab}$. To linear order, the inverse metric is then $g^{ab} = {}^0g^{ab} - \delta g^{ab}$ where

$\delta g^{ab} = {}^0g^{ac}\,{}^0g^{bd}\delta g_{cd}$. Here and in what follows, indices are raised and lowered using the background metric.

4.4.1 Linearized Einstein equation

A perturbed metric provides the perturbed connection $\Gamma^c_{ab} = {}^0\Gamma^c_{ab} + C^c_{ab}$ with the tensor

$$C^c_{ab} = \frac{1}{2}\,{}^0g^{cd}\left({}^0\nabla_a\delta g_{bd} + {}^0\nabla_b\delta g_{ad} - {}^0\nabla_d\delta g_{ab}\right).$$

It determines the Ricci tensor by

$$\begin{aligned} R_{ab} &= {}^0R_{ab} + \partial_c C^c_{ab} - \partial_a C^c_{cb} + {}^0\Gamma^c_{ab}C^d_{cd} + C^c_{ab}\,{}^0\Gamma^d_{cd} \\ &\quad - {}^0\Gamma^c_{db}C^d_{ca} - C^c_{ab}\,{}^0\Gamma^d_{ca} + O(\delta g^2) = {}^0R_{ab} + {}^0\nabla_c C^c_{ab} - {}^0\nabla_a C^c_{cb} + O(\delta g^2) \\ &= {}^0R_{ab} + {}^0\nabla^c\,{}^0\nabla_{(b}\delta g_{a)c} - \frac{1}{2}\,{}^0\nabla^c\,{}^0\nabla_c\delta g_{ab} - \frac{1}{2}\,{}^0\nabla_a\,{}^0\nabla_b\delta g_{cd}\,{}^0g^{cd} + O(\delta g)^2 \end{aligned} \tag{4.86}$$

and the Ricci scalar by

$$R = {}^0R - \delta g^{ab}\,{}^0R_{ab} + {}^0\nabla^a\,{}^0\nabla^b\delta g_{ab} - {}^0\nabla^c\,{}^0\nabla_c\delta g_{ab}\,{}^0g^{ab} + O(\delta g^2).$$

Taken together, this gives the Einstein tensor

$$\begin{aligned} G_{ab} = R_{ab} - \frac{1}{2}Rg_{ab} &= {}^0G_{ab} + {}^0\nabla^c\,{}^0\nabla_{(b}\delta\bar{g}_{a)c} - \frac{1}{2}\,{}^0\nabla^c\,{}^0\nabla_c\delta\bar{g}_{ab} \\ &\quad - \frac{1}{2}\,{}^0\nabla^c\,{}^0\nabla^d\delta\bar{g}_{cd}\,{}^0g_{ab} - \frac{1}{2}\,{}^0R\delta\bar{g}_{ab} + \frac{1}{2}\,{}^0g_{ab}\,{}^0R_{cd}\delta\bar{g}^{cd} \\ &= 8\pi G({}^0T_{ab} + \delta T_{ab}) \end{aligned} \tag{4.87}$$

where we introduce $\delta\bar{g}_{ab} := \delta g_{ab} - \frac{1}{2}\,{}^0g_{ab}\,{}^0g^{cd}\delta g_{cd}$. Since the background solves the unperturbed equation ${}^0G_{ab} = 8\pi G\,{}^0T_{ab}$, for instance by a model of homogeneous cosmology, $\delta\bar{g}_{ab}$ is subject to the partial differential equations

$$\begin{aligned} &{}^0\nabla^c\,{}^0\nabla_c\delta\bar{g}_{ab} + {}^0g_{ab}\,{}^0\nabla^c\,{}^0\nabla^d\delta\bar{g}_{cd} - 2\,{}^0\nabla^c\,{}^0\nabla_{(b}\delta\bar{g}_{a)c} \\ &= {}^0g_{ab}\,{}^0R_{cd}\delta\bar{g}^{cd} - {}^0R\delta\bar{g}_{ab} - 16\pi G\delta T_{ab}. \end{aligned} \tag{4.88}$$

If matter is provided by a field such as a scalar φ, ${}^0T_{ab}$ and δT_{ab} can be obtained by perturbing e.g. $\varphi = {}^0\varphi + \delta\varphi$ and expanding T_{ab} in $\delta\varphi$.

The linearized equations are a system of hyperbolic equations that couple $\delta\bar{g}_{ab}$ to matter and the background curvature. For instance, on Minkowski space as the background ${}^0g_{ab} = \eta_{ab}$, we can gauge δN and N^a to zero and impose other gauge conditions such that only gravitational waves given by the transverse-traceless part of the spatial metric h_{ab} remain. In a cosmological situation, other modes are of interest as well.

4.4.1.1 Mode decomposition

In order to perform the mode decomposition of $\delta g_{\mu\nu}$ on a general background ${}^0g_{\mu\nu}$, we use the canonical form

$$\mathrm{d}x^\mu \mathrm{d}x^\nu \delta g_{\mu\nu} = -2\,{}^0N\delta N \mathrm{d}t^2 + \delta h_{ab}(\mathrm{d}x^a + {}^0N^a \mathrm{d}t)(\mathrm{d}x^b + {}^0N^b \mathrm{d}t) + 2\,{}^0h_{ab}\delta N^a \mathrm{d}t \mathrm{d}x^b$$

for the metric perturbation, where we have perturbed $N = {}^0N + \delta N$, $N^a = {}^0N^a + \delta N^a$ and $h_{ab} = {}^0h_{ab} + \delta h_{ab}$. The coefficients provide metric perturbations of different (spatial) tensorial types. In these perturbations, $\delta N =: {}^0N\phi$ is a spatial scalar and ${}^0h_{ab}\delta N^a$ is a spatial covector which we decompose further as ${}^0h_{ab}\delta N^a = w_b^T + {}^0D_b\sigma_1$ such that ${}^0D^b w_b^T = 0$ (w_a^T is transverse). Thus,

$${}^0D^b\,{}^0D_b\sigma_1 = {}^0D^b({}^0h_{ab}\delta N^a)\,. \tag{4.89}$$

Decompositions of this form allow us to split perturbations into different parts, one coming from a scalar mode, σ_1, another from a vectorial one, w_b^T. Although they both contribute to the spatial vector δN^a, these scalar and vector perturbations obey separate evolution equations at the linear level, and are subject to different sources. The split is thus useful to organize and analyze linear perturbation equations.

In addition to scalar and vector modes of different types, tensor modes arise in a similar way. The spatial tensor δh_{ab} can be decomposed as $\delta h_{ab} = -2\psi\,{}^0h_{ab} + k_{ab}$ where k_{ab} is traceless and $\psi = -\frac{1}{6}\delta h_{ab}\,{}^0h^{ab}$ is a further scalar mode. Moreover, we decompose k_{ab} as

$$\begin{aligned} k_{ab} &= h_{ab}^{TT} + 2\,{}^0D_{(a}u_{b)} - \tfrac{2}{3}\,{}^0h_{ab}\,{}^0D^c u_c \\ &= h_{ab}^{TT} + 2\,{}^0D_{(a}u_{b)}^T + 2({}^0D_a\,{}^0D_b - \tfrac{1}{3}\,{}^0h_{ab}\,{}^0D^c\,{}^0D_c)\sigma_2\,. \end{aligned} \tag{4.90}$$

Here, we are first adding u_b-terms to make h_{ab}^{TT} transverse-traceless, i.e. $h_{ab}^{TT}\,{}^0h^{ab} = 0$ and ${}^0D^a h_{ab}^{TT} = 0$. This requires that ${}^0D^a k_{ab} = 2\,{}^0D^a\,{}^0D_{(a}u_{b)} - \frac{2}{3}\,{}^0D_b\,{}^0D^c u_c$, or

$${}^0D^a\,{}^0D_a u_b + \frac{1}{3}\,{}^0D_b\,{}^0D^a u_a + {}^{0(3)}R_b^d u_d = {}^0D^a k_{ab} \tag{4.91}$$

using the spatial background Ricci tensor ${}^{0(3)}R_{ab}$. In the second step of (4.90), we have treated the co-vector field u_a as we treated w_a before, decomposing it as $u_b = u_b^T + {}^0D_b\sigma_2$ with transverse u_b^T, i.e. ${}^0D^b u_b^T = 0$. For a given u_b solving (4.91), σ_2 is determined as a solution of ${}^0D^b\,{}^0D_b\sigma_2 = {}^0D^b u_b$. The transverse parts are then found as the differences $u_b^T = u_b - {}^0D_b\sigma_2$ and $h_{ab}^{TT} = k_{ab} - 2\,{}^0D_{(a}u_{b)} + \frac{2}{3}\,{}^0h_{ab}\,{}^0D^c u_c$.

For suitable boundary or asymptotic conditions to make the elliptic equations appearing here well-posed, as discussed by Stewart (1990), the original metric perturbations have been split uniquely into the components $(\phi, \psi, \sigma_1, \sigma_2; u_a^T, w_b^T; h_{ab}^{TT})$: four scalars, two vectors and one tensor which constitute ten independent fields. (A transverse vector field has two independent components, since it is subject to one linear transversality condition, while a transverse-traceless symmetric tensor has two independent components.) This parameterizes the ten components of the symmetric space-time metric perturbations.

So far, we have not reduced the amount of information in the variables, but this can be done by gauge conditions which correspond to fixing some of the freedom of making changes of coordinates. These changes are generated by the Lie derivative along arbitrary space-time vector fields ξ^μ such that $x^\mu \mapsto x^\mu + \xi^\mu$. For an FLRW background, which we assume from now on, a coordinate change thus implies a change of the background metric to ${}^0g_{\mu\nu} + \delta_\xi g_{\mu\nu}$ where

$$\begin{aligned} \mathrm{d}x^\mu \mathrm{d}x^\nu \delta_\xi g_{\mu\nu} = &-2\,{}^0N({}^0N\xi^t)^\cdot \mathrm{d}t^2 + 2\left(a^2(a^{-2}\xi_a)^\cdot + {}^0D_a\xi_t\right)\mathrm{d}x^a\mathrm{d}t \\ &+ \left(2a\dot{a}\xi^t\,{}^0h_{ab} + 2\,{}^0D_{(a}\xi_{b)}\right)\mathrm{d}x^a\mathrm{d}x^b\,. \end{aligned} \tag{4.92}$$

This can be computed from the Lie derivative $\delta_\xi g_{\mu\nu} = \mathcal{L}_\xi\,{}^0g_{\mu\nu}$ of the background metric $\mathrm{d}s^2 = -\,{}^0N^2\mathrm{d}t^2 + a^2\,{}^0h_{ab}\mathrm{d}x^a\mathrm{d}x^b$ along ξ^μ, or simply by replacing coordinates x^μ by $x^\mu + \xi^\mu$ in the differentials as well as the arguments of $N(t)$ and $a(t)$. By changing coordinates, we thus transform our original metric perturbation $\delta g_{\mu\nu}$ to $\delta g_{\mu\nu} + \delta_\xi g_{\mu\nu}$ and, by suitable choices of ξ, eliminate some of the components of the perturbation.

For instance, starting with a perturbed metric decomposed as above, we may choose the spatial components of the vector field to be $\xi_a = -u_a$ and the time component ξ_t such that

$${}^0D^a\,{}^0D_a\xi_t + a^2\,{}^0D^a(\xi_a/a^2)^\cdot = -\,{}^0D^a\,{}^0D_a\sigma_1\,. \tag{4.93}$$

Observing that by (4.93) the coordinate change implies a shift change $\delta_\xi N_a = a^2(a^{-2}\xi_a)^\cdot + {}^0D_a\xi_t$ in (4.92) satisfying ${}^0D^a\delta_\xi N_a = -\,{}^0D^a\,{}^0D_a\sigma_1$, we have ${}^0D^a(\delta_\xi N_a + \delta N_a) = 0$ according to the condition (4.89) for σ_1. We thus cancel any contribution σ_1 to the perturbed metric, and similarly eliminate the three other metric components σ_2 and u_a^T because we have chosen $\xi_a = -u_a$. These components are pure gauge: they can be eliminated completely by choosing particular coordinates. Having eliminated four independent fields, all the freedom provided by coordinate transformations has been used. The remaining modes must correspond to physical degrees of freedom, their dynamics determined by evolution equations.

4.4.1.2 *Propagation*

In order to discuss hyperbolicity issues and the physical significance of the modes ϕ, ψ, w_a^T and h_{ab}^{TT} regarding their propagation behavior, we look in more detail at the principal parts of the linearized equations of motion. These are given entirely by the gravitational parts of the equations, and thus independent of the form of matter stress-energy. We read off terms from (4.88) and split them into the perturbed metric components, also using transversality of some modes.

We will use the barred metric perturbation

$$\mathrm{d}x^\mu\mathrm{d}x^\nu\delta\bar{g}_{\mu\nu} = -\,{}^0N^2(\phi+3\psi)\mathrm{d}t^2 + 2w_a^T\mathrm{d}x^a\mathrm{d}t + \left((\psi-\phi)\,{}^0h_{ab} + h_{ab}^{TT}\right)\mathrm{d}x^a\mathrm{d}x^b \tag{4.94}$$

as it follows using $\delta g_{\mu\nu}\,{}^0g^{\mu\nu} = 2(\phi - 3\psi)$. Thanks to transversality of the modes with superscript "T", we have

$${}^0\nabla^\mu\,{}^0\nabla^\nu\delta\bar{g}_{\mu\nu} = {}^0\nabla^t\,{}^0\nabla^t\delta\bar{g}_{tt} + \frac{1}{3}\,{}^0h^{bc}\,{}^0D^a\,{}^0D_a\delta\bar{g}_{bc}\,. \tag{4.95}$$

For the (t, t)-component of Einstein's equation (4.88), we obtain the principal part

$$\begin{aligned} &{}^0\nabla^\mu\, {}^0\nabla_\mu \delta\bar{g}_{tt} + {}^0g_{tt}\, {}^0\nabla^\mu\, {}^0\nabla^\nu \delta\bar{g}_{\mu\nu} - 2\, {}^0\nabla^\mu\, {}^0\nabla_t \delta\bar{g}_{t\mu} \\ &= {}^0D^a\, {}^0D_a \delta\bar{g}_{tt} + {}^0g_{tt}\, {}^0D^a\, {}^0D^b \delta\bar{g}_{ab} = {}^0D^a\, {}^0D_a \left(\delta\bar{g}_{tt} + \tfrac{1}{3}\, {}^0g_{tt}\, {}^0h^{bc}\delta\bar{g}_{bc}\right) \end{aligned} \tag{4.96}$$

using (4.95), transversality of w_a^T in $\delta\bar{g}_{ab}$ and ${}^0g_{tt}\, {}^0\nabla^t = {}^0\nabla_t$ for a diagonal ${}^0g_{ab}$. With $\delta\bar{g}_{tt} + \frac{1}{3}\, {}^0g_{tt}\, {}^0h^{bc}\delta\bar{g}_{bc} = -4\, {}^0N\psi$, (4.96) amounts to an elliptic operator acting on ψ.

For the (a, t)-component with spatial a, we have the principal part

$$\begin{aligned} {}^0\nabla^\mu\, {}^0\nabla_\mu \delta\bar{g}_{t\nu} - {}^0\nabla^\mu\, {}^0\nabla_t \delta\bar{g}_{\mu\nu} - {}^0\nabla^\mu\, {}^0\nabla_\nu \delta\bar{g}_{t\mu} &= {}^0D^b\, {}^0D_b \delta\bar{g}_{ta} - {}^0\nabla^t\, {}^0D_a \delta\bar{g}_{tt} \\ &\quad - {}^0D^b\, {}^0\nabla_t \delta\bar{g}_{ab}\,. \end{aligned} \tag{4.97}$$

The transverse part of this differential operator provides an elliptic equation for w_a^T, the rest a first-order (in time) equation for ${}^0D_a\psi$.

Finally, from the spatial (a, b)-components we have

$${}^0\nabla^\mu\, {}^0\nabla_\mu \delta\bar{g}_{ab} + {}^0h_{ab}\, {}^0\nabla^\mu\, {}^0\nabla^\nu \delta\bar{g}_{\mu\nu} - 2\, {}^0\nabla^\mu\, {}^0\nabla_{(a}\delta\bar{g}_{b)\mu} =: \delta\bar{G}_{ab}\,. \tag{4.98}$$

The tensor $\delta\bar{G}_{ab}$ has a traceless part

$$\delta\bar{G}_{ab} - \frac{1}{3}\, {}^0h^{cd}\delta\bar{G}_{cd}\, {}^0h_{ab} = {}^0\nabla^\mu\, {}^0\nabla_\mu h_{ab}^{TT} - 2\, {}^0\nabla^\mu\, {}^0D_{(a}\delta\bar{g}_{b)\mu} + \frac{2}{3}\, {}^0h_{ab}\, {}^0\nabla^\mu\, {}^0D^d \delta\bar{g}_{\mu d}$$

whose last two terms can be split into its transverse-vector and scalar parts by

$$2\, {}^0\nabla^\mu\, {}^0D_{(a}\delta\bar{g}_{b)\mu} = 2\, {}^0\nabla^t\, {}^0D_{(a}w_{b)}^T + 2\left({}^0D^c\, {}^0D_{(a}\delta\bar{g}_{b)c} - \frac{1}{3}\, {}^0h_{ab}\, {}^0D^c\, {}^0D^c \delta\bar{g}_{cd}\right).$$

Here, the only off-diagonal scalar contribution is ${}^0D^c\, {}^0D_{(a}\delta\bar{g}_{b)c} = {}^0D_a\, {}^0D_b(\psi - \phi)$, providing an elliptic equation for $\psi - \phi$. (If there is no off-diagonal matter source term — called anisotropic stress — this equation is simply solved by $\phi = \psi$.)

The transverse-vector contribution to the trace-free part results in a first-order evolution equation for ${}^0D_a w_b^T$, and the transverse-traceless part of ${}^0\nabla^c\, {}^0\nabla_c \delta\bar{g}_{ab}$ is a hyperbolic operator for h_{ab}^{TT}. Then, the trace-part of $\delta\bar{G}_{ab}$ provides $-4\, {}^0N^2\, {}^0\nabla^t\, {}^0\nabla^t\psi$, which results in a second-order evolution equation for ψ without spatial derivatives. (As expected, we have six equations from the spatial part of Einstein's equation: one for a scalar ψ, one for the difference $\psi - \phi$, two for a transverse vector w_a^T and two for a transverse-traceless symmetric tensor h_{ab}^{TT}.)

4.4.1.3 Equations of motion

From the analysis of principal parts, we expect only h_{ab}^{TT} to be propagating, subject to a wave equation. The other modes are subject only to equations that are either constraints (of less than second order in time derivatives) or purely time derivatives in their principal parts. Such modes evolve, but merely in reaction to sources, not as propagating degrees of freedom. Explicitly, the equations in the presence of a scalar field source φ (whose stress-energy tensor is free of off-diagonal spatial components) as already seen in Chapter 2 are indeed a constraint (2.40) from the (t, a)-equation (corresponding to the diffeomorphism constraint), another constraint (2.39) from the (t, t)-component (corresponding to the Hamiltonian constraint) and an evolution equation (2.41) for the scalar mode. For the vector mode, we have the constraint

$$ {}^0D^b\, {}^0D_b w^T_a = 0 \tag{4.99} $$

and an evolution equation

$$ \frac{\partial}{\partial t}\, {}^0D_{(a} w^T_{b)} = 0 \tag{4.100} $$

(source-free for a scalar matter source), and the wave equation for the tensor mode.

This demonstrates the following properties of the modes:

1. At the linear level, all the modes ϕ, ψ, w^T_a and h^{TT}_{ab} decouple from each other and satisfy separate differential equations.
2. There are no wave equations for the scalar modes ϕ and ψ or the vector mode w^T_a but only (elliptic) constraint equations and evolution equations. Thus, these modes do not propagate but only change in time in reaction to matter sources.
3. For the tensor mode h^{TT}_{ab} we have a wave equation, and thus the transverse-traceless components of the spatial metric, in the gauge chosen here, constitute two propagating degrees of freedom which can be excited even in vacuum: these are gravitational waves.

4.4.2 Gauge-invariant perturbations

In the previous section, we fixed the coordinate gauge by eliminating four of the metric perturbations. Alternatively, one can derive combinations of the general perturbations invariant under changing coordinates by arranging the quantities such that coordinate-dependent terms cancel. These gauge-invariant perturbations are useful because they allow one to do calculations in a coordinate and gauge-independent manner.

4.4.2.1 Space-time treatment

The tensor mode is already gauge invariant because the Lie derivative of the metric along any vector field does not have a transverse-traceless part. Scalar and vector perturbations, however, change under coordinate transformations. A space-time vector field can be decomposed into one transverse vector component and two scalars: the time component and the gradient part of the spatial vector. These contributions provide independent gauge transformations, two scalar and one 2-component transverse vector, corresponding to the four independent coordinate choices.

For vector modes, only coordinate changes are relevant that are generated by a vector field with spatial components satisfying transversality, ${}^0D_a\xi^a = 0$. Using the previous decompositions and (4.92), this implies the transformation $w^{Ta} \mapsto w^{Ta} + a^2\dot{\xi}^a$ and $u^{Ta} \mapsto u^{Ta} + \xi^a$. Thus, in the combination $V^a := w^{Ta} - a^2\dot{u}^{Ta}$ the transformation terms cancel, and V^a is gauge invariant (as well as transverse). This is the coordinate independent measure of the vector mode.

Similarly, one obtains gauge-invariant combinations for scalar modes introduced by Bardeen (1980), the so-called Bardeen variables. (See the articles by Stewart (1990) and Goode (1989) for further properties.) There are four scalar perturbations in a space-time metric: ϕ, ψ, σ_1 and σ_2. It turns out to be useful to redefine some of them by introducing B via $\sigma_1 =: {}^0NB$, and we also rename $\sigma_2 =: E$. If we then choose conformal time η, such that ${}^0N = a$, the scalar contribution to the line element becomes

$$\begin{aligned} \mathrm{d}s^2 = a^2\Big(&-(1+2\phi)\mathrm{d}\eta^2 + 2\partial_a B\mathrm{d}\eta\mathrm{d}x^a \\ &+ \big((1-2\psi)\delta_{ab} + 2(\partial_a\partial_b - \tfrac{1}{3}\delta_{ab}\partial^c\partial_c)E\big)\,\mathrm{d}x^a\mathrm{d}x^b\Big) \end{aligned} \tag{4.101}$$

with ordinary partial derivatives ∂_a on the spatially flat background (which differ from ${}^0D_a$ by a factor of a). Conformal-time derivatives will be denoted by primes.

A gauge transformation, given by the Lie derivative along a space-time vector field ξ^μ, now gives rise to two scalar parameters ξ^0 and the gradient part ξ such that $\xi_a = \partial_a\xi$. Two independent gauge transformations for scalar modes arise:

- If only ξ_0 is non-zero, we change η to $\eta + \xi^0$, thus $\mathrm{d}\eta^2$ becomes, to first order in ξ^0,

  $$\mathrm{d}\eta^2 + 2\xi^{0\prime}\mathrm{d}\eta^2 + 2\partial_a\xi^0\mathrm{d}\eta\mathrm{d}x^a$$

  and $a(\eta)^2$ changes to $a(\eta)^2(1+2a'\xi^0/a)$. From this, we read off the transformation formulas

  $$\phi \mapsto \phi + \xi^{0\prime} + \frac{a'}{a}\xi^0\,, \qquad \psi \mapsto \psi - \frac{a'}{a}\xi^0\,, \qquad B \mapsto B - \xi^0\,, \qquad E \mapsto E\,. \tag{4.102}$$

- If only ξ is non-zero, $\delta_{ab}\mathrm{d}x^a\mathrm{d}x^b$ changes to

  $$\delta_{ab}\mathrm{d}x^a\mathrm{d}x^b + 2\partial_a\xi'\mathrm{d}\eta\mathrm{d}x^a + 2\partial_a\partial_b\xi\mathrm{d}x^a\mathrm{d}x^b$$

  from which we read off

  $$\phi \mapsto \phi\,, \qquad \psi \mapsto \psi\,, \qquad B \mapsto B + \xi'\,, \qquad E \mapsto E + \xi\,. \tag{4.103}$$

We deduce that $B - E'$ is invariant under ξ-transformations and changes to $B - E' - \xi^0$ under ξ^0-transformations. Thus,

$$\Phi := \phi + \frac{a'}{a}(B - E') + (B - E')' \quad \text{and} \quad \Psi := \psi - \frac{a'}{a}(B - E') \tag{4.104}$$

are gauge invariant.

When there is an additional field due to matter, it also changes under coordinate transformations. This can be used to eliminate the two scalar perturbations B and E from the expressions of gauge-invariant perturbations. A scalar field φ only changes under ξ^0 by $\delta\varphi + {}^0\varphi'\xi^0$ where $\delta\varphi$ is the linear perturbation of the field around its homogeneous value ${}^0\varphi$. Thus,

$$\delta\varphi + {}^0\varphi'(B - E') \tag{4.105}$$

is gauge invariant. This allows us to eliminate $B - E'$ in (4.104) and obtain the perturbations

$$\mathcal{R}_1 = \psi + \frac{a'}{a\,{}^0\varphi'}\delta\varphi$$
$$\mathcal{R}_2 = \phi - \frac{1}{2}\left(\frac{a}{a'}\right)'\psi - \frac{1}{{}^0\varphi'}\left(\frac{a'}{a} - \frac{{}^0\varphi''}{{}^0\varphi'}\right)\delta\varphi + \frac{1}{2}\frac{a}{a'}\psi' - \frac{1}{2\,{}^0\varphi'}\delta\varphi' \qquad (4.106)$$

as gauge invariant quantities independent of B and E. The so-called *curvature perturbation* $\mathcal{R}_1$ will be used later in the context of inflationary structure formation.

4.4.2.2 Canonical treatment

It is instructive to derive the gauge-invariant perturbations from a canonical perspective, done without reference to the space-time manifold or coordinates. One interesting issue is how the gauge transformations for time-time and time-space components of the metric arise. (A Hamiltonian formulation of cosmological perturbations has been provided by Langlois (1994), but using some information about gauge-invariant variables known from the covariant treatment.)

Scalar modes To start with, we have to specify the change of gauge by the form of lapse and shift, which corresponds to the form of the initial metric in a covariant treatment. We do so by introducing the gauge parameters

$$\epsilon^0 = \delta N = {}^0N\xi^0, \quad \epsilon^a = \delta N^a = \xi^a = \partial^a\xi,$$

according to (3.96), with a free background value ${}^0N$, which would be ${}^0N = 1$ for proper time and ${}^0N = a$ for conformal time. (Since we are perturbing around FLRW models, there is no background shift vector: ${}^0N^a = 0$.)

Since some of the metric perturbations are those of lapse and shift, we have to use the general gauge generator (3.93) on the extended phase space introduced in Chapter 3.3.4.3:

$$G[\epsilon^\mu, \dot\epsilon^\mu] = C[\epsilon^\mu] + p[\dot\epsilon^\mu + \langle N, \epsilon\rangle^\mu]$$

with

$$\langle N, \epsilon\rangle^\mu = \int \mathrm{d}^3x\mathrm{d}^3y N^\nu(x)\epsilon^\lambda(y)F^\mu_{\nu\lambda}(x, y; z)$$

in terms of the structure functions $F^\mu_{\nu\lambda}(x, y; z)$ of gravity.

We first look at the transformations of lapse and shift perturbations in $N = {}^0N(1 + \phi)$ and $N^a = {}^0N\partial^a B$, generated by the $p[\cdot]$-part of the gauge generator $G[\epsilon^\mu, \dot\epsilon^\mu]$. These transformations depend only on the constraint algebra via the structure constants, but not on the explicit form of the constraints. For the transformation of ϕ we need the 0-component of $\langle N, \epsilon\rangle^\mu$, which is the contribution producing a Hamiltonian constraint in the constraint algebra. Such a term can only come from the Poisson bracket of a Hamiltonian with a diffeomorphism constraint, for which the general hypersurface deformation algebra provides $\langle N, \epsilon\rangle^0 = \mathcal{L}_{N^a}\epsilon^0 - \mathcal{L}_{\epsilon^a}N^0$. Here, ϵ^μ only has first-order terms while N^a has a vanishing background value and ${}^0N$ is spatially constant. There are thus no zeroth or first-order terms in

$\langle N, \epsilon\rangle^0$, and it does not contribute to gauge transformations of linear perturbations. The only linear gauge transformation of δN comes from $\dot{\epsilon}^0$ in $G[\epsilon^\mu, \dot{\epsilon}^\mu]$: $\delta_{[\epsilon^0,\epsilon]}(\delta N) = \dot{\epsilon}^0 = ({}^0N\xi^0)^{\cdot}$. By definition of ϕ, this equals ${}^0N\delta_{[\epsilon^0,\epsilon]}\phi = {}^0N(\dot{\xi}^0 + {}^0\dot{N}\xi^0/{}^0N)$. (In conformal time, this means that $\delta_{[\epsilon^0,\epsilon]}\phi = \xi^{0\prime} + a'\xi^0/a$ in agreement with the result (4.102) obtained from a Lie derivative.) By analogy, we obtain the spatial part of $\langle N, \epsilon\rangle^a$ from the Poisson brackets of two diffeomorphism or two Hamiltonian constraints. From two diffeomorphism constraints we do not produce a non-zero linear result, while two Hamiltonian constraints give us $\langle N, \epsilon\rangle^a = -{}^0h^{ab}N\partial_b\epsilon^0 = -{}^0N\partial^a\epsilon^0$. Together with the term from $\dot{\epsilon}^a$ and the definition of the metric perturbation B according to $\delta N^a = {}^0N\partial^a B$, we have $\delta_{[\epsilon^0,\epsilon]}B = -\xi^0 + \dot{\xi}$ independently of the choice of background lapse.

Gauge transformations of perturbed dynamical fields are generated by the secondary constraints, whose form now plays a larger role. For gauge transformations of linear variables, we need gauge generators expanded to second order (counting also the order of the Lagrange multipliers lapse and shift):

$$\delta_{[\epsilon^0,\epsilon]}X \equiv \{X, H^{(2)}[{}^0N\xi^0]\} + \{X, D^{(2)}[\partial^a\xi]\}\,.$$

Poisson brackets of linear perturbations with second-order constraints will then produce linear results. Since our multipliers ${}^0N\xi^0$ and $\partial^a\xi$ used to generate gauge transformations are already of first order, the phase-space expressions of the constraints need be expanded only linearly. Following Bojowald et al. (2009b), we will do the explicit calculations in canonical variables of Example 3.21 given by a densitized triad E^a_i and extrinsic curvature K^i_a, which is somewhat in between the ADM and Ashtekar–Barbero variables.

We thus perturb the densitized triad $E^a_i = {}^0E^a_i + \delta E^a_i$ with

$${}^0E^a_i = \tilde{p}\delta^a_i\,, \qquad \delta E^a_i = -2\tilde{p}\psi\delta^a_i + p(\delta^a_i\Delta - \partial^a\partial_i)E \tag{4.107}$$

for scalar modes and with the background triad parameter $|\tilde{p}| = a^2$. Similarly, $K^i_a = \tilde{k}\delta^i_a + \delta K^i_a$ with expressions for $\tilde{k}$ and δK^i_a in terms of time derivatives of the triad components derived later from the equations of motion. For the constraints expanded to second order, one may use the linear expression

$$\delta\Gamma^i_a = \frac{1}{\tilde{p}}\epsilon^{ije}\delta_{ac}\partial_e\delta E^c_j \tag{4.108}$$

for the spin connection (whose background value vanishes in the case of a spatially flat model considered here), and we find the Hamiltonian constraint

$$H^{(2)}[\delta N] = \frac{1}{16\pi G}\int \mathrm{d}^3x\,\delta N\left(-4\tilde{k}\sqrt{|\tilde{p}|}\delta^c_j\delta K^j_c - \frac{\tilde{k}^2}{\sqrt{|\tilde{p}|}}\delta^j_c\delta E^c_j + \frac{2}{\sqrt{|\tilde{p}|}}\partial_c\partial^j\delta E^c_j\right)\,. \tag{4.109}$$

Matter fields can be included similarly by expansions. The diffeomorphism constraint has the second-order term

$$D^{(2)}[\delta N^a] = \frac{1}{8\pi G}\int_\Sigma \mathrm{d}^3x\,\delta N^a\left(\tilde{p}\partial_a(\delta^d_k\delta K^k_d) - \tilde{p}(\partial_k\delta K^k_a) - \tilde{k}\delta^k_a(\partial_d\delta E^d_k)\right)\,. \tag{4.110}$$

Gauge transformations of some variables will depend on k, which we express in terms of a using the background constraints. Only the background Hamiltonian constraint is non-vanishing, and its gravitational part, which is sufficient for now, reads

$$H^{(0)}_{\rm grav}[{}^0N] = -\frac{3V_0}{8\pi G}\,{}^0N\sqrt{|\tilde{p}|}\tilde{k}^2\,. \tag{4.111}$$

One of the equations of motion, $\dot{\tilde{p}} = \{\tilde{p}, H^{(0)}_{\rm grav}[{}^0N]\} = 2\,{}^0N\sqrt{|\tilde{p}|}\tilde{k}$, thus tells us that $\tilde{k} = \dot{\tilde{p}}/(2\,{}^0N\sqrt{|\tilde{p}|})$ (which in conformal time specializes to the conformal Hubble parameter $\tilde{k} = \tilde{p}'/2\tilde{p} = a'/a = \mathcal{H}_{\rm conf}$).

With the expanded constraints, the spatial part of the metric, encoded in the densitized triad, transforms as

$$\delta_{[\epsilon^0,\epsilon]}\delta E^a_i = 2\tilde{k}\sqrt{|\tilde{p}|}\epsilon^0\delta^a_i + \tilde{p}(\delta^a_i\Delta\epsilon - \partial^a\partial_i\epsilon) \tag{4.112}$$

under gauge transformations. By comparing this with the general expansion (4.107) of the densitized triad, we see

$$\delta_{[\epsilon^0,\epsilon]}\psi = -\frac{\tilde{k}}{\sqrt{|\tilde{p}|}}\epsilon^0 = -\frac{\dot{a}}{a}\xi^0\,, \qquad \delta_{[\epsilon^0,\epsilon]}E = \epsilon = \xi\,. \tag{4.113}$$

To summarize, we have the gauge transformations (4.113) together with

$$\delta_{[\epsilon^0,\epsilon]}\phi = \dot{\xi}^0 + \frac{{}^0\dot{N}}{{}^0N}\xi^0\,, \qquad \delta_{[\epsilon^0,\epsilon]}B = -\xi^0 + \dot{\xi} \tag{4.114}$$

in an arbitrary background gauge specified by ${}^0N$. The two gauge-invariant scalar combinations are

$$\Phi := \phi + \frac{{}^0\dot{N}}{{}^0N}(B - \dot{E}) + (B - \dot{E})^\bullet \quad \text{and} \quad \Psi := \psi - \frac{\dot{a}}{a}(B - \dot{E})\,. \tag{4.115}$$

Vector modes The only other gauge transformation is generated by the transverse part of ϵ^a, acting on the vector mode. Among the non-dynamical fields, only the shift vector is involved in vector modes, and since linear-order terms in $\langle N, \epsilon\rangle^a$ are all gradients but not transverse, the vector mode in $\delta N^a =: S^a$ receives gauge transformations only from the $\dot{\epsilon}^a$-part in (3.93): $\delta_{[0,\epsilon^{Ta}]}S^a = \dot{\epsilon}^{Ta}$.

The spatial part of the metric contains a vector mode via u^T_b in k_{ab} of (4.90). Defining $u^T_b =: a^2F_b$, it produces a linearly perturbed densitized triad

$$E^a_i = \tilde{p}\delta^a_i - \partial^a u^T_i = \tilde{p}(\delta^a_i - \partial_i F^a)\,.$$

Computing the corresponding metric (noting that the determinant of the perturbed metric does not have linear terms thanks to the transversality of F^a), we indeed produce the correct form $h_{ab} = a^2(\delta_{ab} + \partial_a F_b + \partial_b F_a)$.

For gauge transformations of the spatial-metric perturbations, we again use the second-order constraints. The Hamiltonian constraint (4.109) does not contribute to gauge transformations of δE^a_i, since only the non-transverse momentum $\delta^a_k\delta K^k_a$ appears in it. From the diffeomorphism constraint, we generate $\delta_{[0,\epsilon^{Ta}]}(\delta E^a_i) = -\tilde{p}\partial_i\epsilon^{Ta}$, or

$$\delta_{[0,\epsilon^{Ta}]}F^a = \epsilon^{Ta}\,, \qquad \delta_{[0,\epsilon^{Ta}]}S^a = \dot{\epsilon}^{Ta}\,. \tag{4.116}$$

The gauge-invariant combination of vector modes is $\sigma^a := S^a - \dot{F}^a$.

4.4.3 Gauge-invariant equations of motion

Covariantly, cosmological perturbation equations are obtained by linearizing the Einstein and stress-energy tensors around an FLRW solution. Although the equations are over-determined, they are fully consistent and can be expressed purely in terms of the gauge-invariant perturbations. Canonically, all the required equations follow from the constraints, and consistency as well as gauge independence is ensured by the first-class nature of the constraints: first-class constraint equations $C_I = 0$ are automatically preserved by the evolution equations they generate, $\dot{C}_I = \{C_I, \sum_J N^J C_J\} \approx 0$ in a choice of frame specified by the multipliers N^I. One can demonstrate these aspects by examples for the canonical perturbation equations of general relativity.

As before, we need the constraints expanded to second order to generate equations of motion for linear modes. However, unlike in the case of gauge transformations, the multipliers inserted in the constraints now have zeroth-order terms when we generate evolution: the background frame is specified by a non-vanishing lapse ${}^0N$ (while the background shift still vanishes for perturbations around an isotropic model). We can use the second-order diffeomorphism constraint as obtained before, but the second-order Hamiltonian constraint acquires new terms quadratic in δE^a_i and δK^i_a, and multiplied with ${}^0N$. We will not present this more lengthy constraint here and end the discussion of equations of motion with two examples.

Example 4.18 (Diffeomorphism constraint)
The second-order diffeomorphism constraint implies a linear equation by variation with respect to δN^a. Including the matter part for a scalar field φ, which is $\int_\Sigma \mathrm{d}^3x\delta N^c\,{}^0p_\varphi\partial_c\delta\varphi$, the equation for linear metric perturbations is

$$\tilde{p}\left(\partial_c(\delta K^i_a\delta^a_i) - \partial_k(\delta K^k_c)\right) - \tilde{k}\partial_d\delta E^d_k\delta^k_c + 8\pi G\,{}^0p_\varphi\partial_c\delta\varphi = 0\,.$$

From the equation of motion for δE^a_i we obtain an equation expressing δK^i_a in terms of the metric perturbations and their time derivatives. Using the Bardeen variables, this is

$$\delta K^i_a = -\delta^i_a\left(\Psi' + \mathcal{H}_{\mathrm{conf}}(\Psi+\Phi)\right) - \delta^i_a\mathcal{H}'_{\mathrm{conf}}(B-E') + \partial_a\partial^i\left(\mathcal{H}_{\mathrm{conf}}E - (B-E')\right).$$

Inserting this expression in the diffeomorphism constraint results in

$$\partial_c\left(\Psi' + \mathcal{H}_{\mathrm{conf}}\Phi - 4\pi G\,{}^0\varphi'\delta\varphi^{\mathrm{GI}}\right) = 0\,, \tag{4.117}$$

producing Eq. (2.40).

In a similar way, the Hamiltonian constraint equation and the evolution equations can be expressed solely in terms of gauge-invariant quantities, owing to the first-class nature of the constraints.

Example 4.19 (Tensor mode)
For tensor modes, we perturb the densitized triad as

$$E^a_i = \tilde{p}\left(\delta^a_i - \frac{1}{2}h^a_i\right) \tag{4.118}$$

with the tracelessness condition $\delta^i_a \delta E^a_i = 0$ *as well as transversality. Computing the spatial metric, this can easily be seen to result in a metric perturbation* $h^{TT}_{ab} = \delta_{i(b}h^i_{a)}$*. If we combine* $\delta \dot{E}^a_i = -\frac{1}{2}(\tilde{p}\dot{h}^a_i + \dot{\tilde{p}}h^a_i)$ *with the canonical equation of motion for* E^a_i*, we obtain* $(h^i_a)'' + 2\mathcal{H}_{\rm conf}(h^i_a)' - {}^0D^2 h^i_a = 0$ *in conformal time.*

For some purposes in modern cosmology, higher-order perturbations starting with the second order must be used. Techniques to organize these calculations have been developed, e.g., by Malik and Wands (2004) and Brizuela et al. (2007). Gauge-invariant quantities defined at all orders (and even frame-independent) have been introduced by Ellis and Bruni (1989); Bruni et al. (1992) and used by Langlois and Vernizzi (2005). One of the results is a general statement about the evolution of large-scale modes, which we will discuss using canonical methods in the next section.

4.4.4 Basics of inflationary structure formation

According to the inflationary scenario, all the large-scale structures one sees in the universe, such as anisotropies of the cosmic microwave radiation or the galaxy distribution, can be phenomenologically described as arising from vacuum fluctuations of a single scalar field φ, blown out of proportion by an early accelerating phase of the universe.

During inflation, we have $\ddot{a} > 0$ for a universe nearly isotropic on large scales. The Hubble radius $1/a\mathcal{H} = 1/\dot{a}$ is decreasing, and if one waits for a sufficiently long time, any inhomogeneity mode (of metric perturbations or the scalar field) of wave number k will at some point "exit the Hubble radius", i.e. the Hubble radius will become less than the mode's wavelength. As we will see in detail, the dynamical behavior of modes outside the Hubble radius is quite special and, when put into the right form, follows simple evolution equations governed by conservation laws. Irrespective of the precise expansion of the background, which very early on in the universe would depend on largely unknown equations of state of exotic matter forms, large-scale modes can easily be followed from the very early time when they leave the Hubble radius to a much later time, long after inflation has ended, when the decelerated radiation or matter-dominated expansion has grown the Hubble radius back to a size larger than the wavelength of the mode considered. We are thus able to bridge a long time-span from suitable initial conditions at the onset of inflation all the way to the seeds of structure in a universe filled with ordinary matter or radiation.

4.4.4.1 Conservation of power on large scales

Linear perturbations are sufficient to discuss most of these basic aspects quantitatively. But the behavior of modes outside the Hubble radius can be seen in a more general, non-linear way using the so-called separate-universe approach of Salopek and Bond (1990); Wands et al. (2000) and Bertschinger (2006). Here, one employs the full canonical equations for the spatial metric and extrinsic curvature, but drops all second-order spatial derivatives (which are subdominant for modes of long wavelength compared to time derivatives or factors of $a\mathcal{H}$), and sets $N^a = 0$. Although non-linear inhomogeneities are present, the resulting equations will look much like those of an isotropic universe.

Equations of motion First, in a way very similar to the BSSN scheme in Chapter 3.4.2.3, we split the extrinsic-curvature tensor K_{ab} into its trace part, the expansion $\theta := h^{ab}K_{ab}$ which can be used as a measure for the local Hubble parameter $\mathcal{H} = \theta/3$, and the trace-free shear term $\sigma_{ab} := K_{ab} - \frac{1}{3}\theta h_{ab}$. A conformal decomposition $h_{ab} =: h^{1/3}\tilde{h}_{ab}$ with $h := \det h_{ab}$ then results in

$$\theta = \frac{1}{2Nh}\frac{\partial h}{\partial t}\,, \qquad \sigma_{ab} = \frac{1}{2N}h^{1/3}\frac{\partial \tilde{h}_{ab}}{\partial t} \tag{4.119}$$

computed using (3.40) to arrive at θ, then using (3.51) with a vanishing shift vector.

Dropping all terms containing second-order spatial derivatives as well as those containing the shift vector results in simplified constraints

$$\sigma_{ab}\sigma^{ab} - \frac{2}{3}\theta^2 + 16\pi G\rho_{\mathrm{E}} = 0 \tag{4.120}$$

$$-D_b\sigma_a^b + \frac{2}{3}D_a\theta - 8\pi G J_a^{\mathrm{E}} = 0 \tag{4.121}$$

from (3.63) and (3.64) using $p_{ab} = (16\pi G)^{-1}\sqrt{h}(\sigma_{ab} - \frac{2}{3}\theta h_{ab})$. Matter terms provide the energy density $\rho_{\mathrm{E}} = N^{-2}T_{00}$ (Eq. (3.171) with $N^a = 0$) and the energy current $J_a^{\mathrm{E}} = -N^{-1}T_{0a}$. Equations of motion for K_{ab} from (3.80) are

$$\frac{1}{N}\frac{\partial\theta}{\partial t} = -\frac{1}{3}\theta^2 - \sigma_{ab}\sigma^{ab} - 4\pi G N(\rho_{\mathrm{E}} + 3P_{\mathrm{E}}) \tag{4.122}$$

$$\frac{1}{N}\frac{\partial\sigma_b^a}{\partial t} = -\theta\sigma_b^a - 8\pi G\tilde{S}_{\mathrm{E}b}^a \tag{4.123}$$

where $P_{\mathrm{E}} = \frac{1}{3}h^{ab}S_{\mathrm{E}ab}$ is pressure and the trace-free $\tilde{S}_{\mathrm{E}ab} = S_{\mathrm{E}ab} - P_{\mathrm{E}}h_{ab}$ anisotropic stress. (See also Exercise 4.2.)

In the absence of anisotropic stress, the shear, combining (4.119) and (4.123), must satisfy the equation

$$\frac{\partial\sigma_b^a}{\partial t} = -\frac{\sigma_b^a}{2h}\frac{\partial h}{\partial t}$$

which implies that $\sigma_b^a = (h/h_0)^{-1/2}(\sigma_0)_b^a$ in terms of an initial shear $(\sigma_0)_b^a$ decays as the universe expands. Terms containing σ_b^a can thus be dropped, simplifying the equations further. We are left with the Hamiltonian constraint $\frac{1}{3}\theta^2 = 8\pi G\rho_{\rm E}$, the diffeomorphism constraint

$$D_a\theta = 12\pi G J_a^{\rm E} \tag{4.124}$$

and the evolution equation

$$\frac{1}{3N}\frac{\partial\theta}{\partial t} = -4\pi G(\rho_{\rm E} + P_{\rm E}) \tag{4.125}$$

from (4.122), using the Hamiltonian constraint. The first one of these equations is equivalent to the Friedmann equation of an isotropic, spatially flat universe with Hubble parameter $\mathcal{H} = \frac{1}{3}\theta$, and the third one amounts to the Raychaudhuri equation. At this stage, large-scale modes of inhomogeneities are seen to follow the FLRW dynamics. Different spatial regions follow this dynamics independently, except that initial conditions must obey the constraint equation (4.124).

Consistency We have obtained a simplified system of equations by assuming spatial derivatives to be small. As zeroth-order quantities in this kind of derivative expansion, we have θ, N and the matter parameters $\rho_{\rm E}$ and $P_{\rm E}$. Of first order are the energy current $J_a^{\rm E}$ (which for matter fields depends on spatial derivatives of the field) and explicit spatial derivatives such as $\partial_a\theta$. Although we have started with a consistent set of constraint and evolution equations, consistency after the truncation at first-derivative order is not automatically guaranteed, but it can be demonstrated explicitly in this case based on the conservation equation of matter.

The Hamiltonian constraint together with the evolution equation, although now formulated for spatially varying fields θ, $\rho_{\rm E}$ and $P_{\rm E}$, just provides us with the system of equations realized in isotropic cosmology, which is known to be consistent. The crucial new equation in the separate-universe approximation is the diffeomorphism constraint (4.124), restricting the spatial dependence of the fields. Taking a time derivative of (4.124) and a spatial derivative of (4.125) produces two equations for $\partial_t\partial_a\theta$ which must agree for consistency. From the matter terms, this requires a relationship between the time derivative of the energy current and spatial derivatives of energy density and pressure as it is provided by stress-energy conservation.

Example 4.20 (Consistency)
We compute the conservation equation $0 = \nabla_\mu T_a^\mu = \partial_\mu T_a^\mu + \Gamma^\mu_{\mu\nu} T_a^\nu - \Gamma^\nu_{\mu a} T_\nu^\mu$ *for a spatial index a, assuming a vanishing shift vector and vanishing anisotropic stress* $\tilde{S}_{{\rm E}ab}$*. Stress-energy components can then be identified as* $T_0^0 = g^{00}T_{00} = -\rho_{\rm E}$, $T_a^0 = g^{00}T_{0a} = N^{-1}J_a^{\rm E}$ *and* $T_a^b = g^{bc}T_{ca} = S_{{\rm E}a}^b = P_{\rm E}h_a^b$.

Partial derivatives of T_a^0 in the conservation equation then are $\partial_\mu T_a^\mu = N^{-1}\partial_t J_a^{\rm E} + \partial_a P_{\rm E} - \dot{N}N^{-2}J_a^{\rm E}$. The Christoffel terms, making use of the vanishing shift, are

$$\begin{aligned}\Gamma^\mu_{\mu\nu}T_a^\nu &= \frac{1}{2}\left(g^{00}\partial_t g_{00} + \dot{h}/h\right)T_a^0 + \frac{1}{2}\left(g^{00}\partial_b g_{00} + \frac{\partial_b h}{h}\right)T_a^b\\ &= \frac{\dot{N}}{N^2}J_a^{\rm E} + \theta J_a^{\rm E} + \left(\frac{\partial_a N}{N} + \frac{\partial_a h}{h}\right)P_{\rm E}\\ \Gamma^\nu_{\mu a}T^\mu_\nu &= \frac{1}{2}g^{00}\partial_a g_{00}T_0^0 - \frac{1}{2}g^{00}\partial_t g_{ba}T_0^b + \frac{1}{2}g^{bc}\partial_t g_{ac}T_b^0 + \frac{1}{2}\frac{\partial_b}{h}T_a^b\\ &= -\frac{\partial_a N}{N}\rho_{\rm E} + \frac{1}{2}\frac{\partial_a h}{h}P_{\rm E}\,.\end{aligned}$$

In the last step, two terms have cancelled thanks to $g^{bc}T_c^0 = T^{0b} = g^{c0}T_c^b = g^{00}T_0^b$ when the shift vector vanishes.

Combining all the terms results in $\partial_t J_a^{\rm E} = -N\theta J_a^{\rm E} - N\partial_a P_{\rm E} - (\rho_{\rm E} + P_{\rm E})\partial_a N$ as part of the conservation equation of the stress-energy tensor. With this equation (4.124) is preserved by the equation of motion (4.125): from the Friedmann equation and (4.124), we first obtain $\frac{2}{3}\theta\partial_a\theta = 8\pi G\partial_a\rho_{\rm E} = 8\pi G\theta J_a^{\rm E}$, and thus $\partial_a\rho_{\rm E} = J_a^{\rm E}\theta$. Then, the time derivative $\partial_t C_a = \partial_a\partial_t\theta - 12\pi G\partial_t J_a^{\rm E} = -12\pi G(\partial_a(N(\rho_{\rm E} + P_{\rm E})) + \partial_t J_a^{\rm E})$ of the diffeomorphism constraint $C_a = \partial_a\theta - 12\pi G J_a^{\rm E}$ automatically vanishes; the large-scale approximation is consistent.

For a specific form of matter, consistency of the equations implies a conservation law for a certain combination of metric components and matter fields. We derive this for matter provided by a scalar φ, such that $\partial_a\mathcal{H} = -4\pi G\Pi\partial_a\varphi$ and $\partial_t\mathcal{H} = -4\pi GN\Pi^2$ with $\Pi := \dot{\varphi}/N$ (which is not the scalar momentum, since it lacks a factor of a^3). The Friedmann equation will not be used for the present argument. Taking a time derivative of the constraint equation and a spatial derivative of the evolution equation, we have

$$\begin{aligned}\partial_t\partial_a\mathcal{H} &= -4\pi G(\dot{\Pi}\partial_a\varphi + \Pi\partial_a\dot{\varphi})\\ &= \partial_a\partial_t\mathcal{H} = -8\pi GN\Pi\partial_a\Pi - 4\pi G\Pi^2\partial_a N = -8\pi G\Pi\partial_a\dot{\varphi} + 4\pi G\Pi^2\partial_a N\end{aligned}$$

and thus $\Pi\partial_a\dot{\varphi} = \dot{\Pi}\partial_a\varphi + \Pi^2\partial_a N$. If we now take a time derivative of the curvature perturbation

$$\mathcal{R}_a := \frac{1}{6}\frac{\partial_a h}{h} - \frac{\mathcal{H}}{\Pi}\partial_a\varphi \tag{4.126}$$

with the local Hubble parameter $\mathcal{H} = \frac{1}{6}\dot{h}/Nh$, we have

$$\frac{1}{N}\partial_t\mathcal{R}_a = \partial_a\mathcal{H} + \mathcal{H}\frac{\partial_a N}{N} - \frac{\dot{\mathcal{H}}}{N\Pi}\partial_a\varphi + \frac{\mathcal{H}\dot{\Pi}}{N\Pi^2}\partial_a\varphi - \frac{\mathcal{H}}{N\Pi}\partial_a\dot{\varphi} = 0 \tag{4.127}$$

and $\mathcal{R}_a$ is conserved on a large scale. (Note that $\partial_a(\dot{h}/h) = \partial_t(\partial_a h/h)$.) For linear perturbations, $\mathcal{R}_a = \partial_a\mathcal{R}_1$ in terms of the curvature perturbation defined earlier, Eq. (4.106).

4.4.4.2 Initial conditions

The curvature perturbation that becomes conserved on super-Hubble scales obeys a simple evolution equation in a space-time gauge in which linear metric perturbations vanish. For its behavior, the Klein–Gordon equation controlling $\delta\varphi$ is sufficient. If we are able to pose reasonable initial conditions in the asymptotic past, the values obtained at Hubble exit can then be computed by integrating the evolution. According to the inflationary scenario, a suitable asymptotic initial condition is given by quantum fluctuations in the vacuum state, as developed by Lyth (1985) and Guth and Pi (1985) as well as, with special emphasis on the classical–quantum transition, by Polarski and Starobinsky (1996); Kiefer et al. (1998). Reviews are available from Lidsey et al. (1997), Bassett et al. (2006) and Baumann (2007).

Before a Fourier-mode $\delta\varphi_k$ of wave number k leaves the Hubble radius, i.e. before k becomes smaller than $a\mathcal{H}$, it follows inflationary dynamics. In this potential-dominated regime of the scalar one assumes that the potential is flat enough for the kinetic energy of the background scalar to remain small. Derivatives of the potential must be small, which can systematically be formulated in the slow-roll approximation, an expansion in terms of normalized derivatives of the potential.

Evolution of Fourier modes For the Fourier modes $\delta\varphi_k$, this means that we can ignore the potential term $\mathrm{d}^2V/\mathrm{d}\varphi^2\delta\varphi_k$ compared to the others in the Klein–Gordon equation. Before a mode leaves the Hubble radius, its dynamics is thus determined by

$$\delta\varphi_k'' + 2\mathcal{H}\delta\varphi_k' + k^2\delta\varphi_k = 0$$

or, removing the first-order term by substituting $v_k := a\delta\varphi_k$,

$$v_k'' + (k^2 - a''/a)v_k = 0\,. \tag{4.128}$$

For near-de Sitter expansion, $a''/a = 2/\eta^2$ in terms of conformal time η (see Example 2.1), which makes the equation solvable:

$$v_k(\eta) = Ae^{-ik\eta}(1 - i/k\eta) + Be^{ik\eta}(1 + i/k\eta) \tag{4.129}$$

with two constants A and B determined via the initial values.

Equation (4.128), derived independently by Mukhanov and Sasaki, has the form of a single harmonic-oscillator equation for each mode, with a time-dependent frequency $\sqrt{k^2 - 2/\eta^2}$. In the asymptotic past, the time dependence vanishes and the frequency is just the wave number k. With this observation, we can perform a heuristic quantization that provides initial values for v_k as they would be realized by a single harmonic oscillator starting its evolution in the ground state. Expectation values of position, representing v_k here, and momentum vanish in the ground state, not giving rise to non-trivial evolution. But fluctuations cannot vanish, and so one may take the quantum fluctuation Δv_k as a reasonable measure for the initial size of the perturbation.

Quantum fluctuations The correct size of quantum fluctuations can only be computed from an action or Hamiltonian, since a classical equation of motion, even if it could be quantized sidestepping actions or Hamiltonians, does not provide a normalization for the v_k. An analysis of the canonical structure underlying the equation of motion provides a unique normalization of quantum fluctuations, provided that the state the modes are in is specified.

We have obtained the equation of motion for v_k in a space-time gauge with vanishing metric perturbations; thus, only the scalar field is inhomogeneous. With an isotropic metric background, the action $S_{\mathrm{scalar}}[\varphi] = -\frac{1}{2}\int \mathrm{d}^4x\sqrt{-\det g}\,g^{\mu\nu}(\partial_\mu\varphi)(\partial_\nu\varphi)$ (in the absence of a potential) reduces to $\frac{1}{2}\int \mathrm{d}\eta\left((a\varphi')^2 - \delta^{ab}(\partial_a\varphi)(\partial_b\varphi)\right)$ in conformal time η. Using $v = a\varphi$ as before, we can write this action as

$$\frac{1}{2}\int \mathrm{d}\eta\left((v')^2 - 2\mathcal{H}_{\mathrm{conf}}vv' + \mathcal{H}_{\mathrm{conf}}^2v^2 - h^{ab}(\partial_a v)(\partial_b v)\right).$$

Integrating by parts in the second term allows us to identify $\int \mathrm{d}\eta\mathcal{H}_{\mathrm{conf}}vv' = -\frac{1}{2}\int \mathrm{d}\eta\mathcal{H}'_{\mathrm{conf}}v^2$ such that the Lagrangian density becomes

$$\mathcal{L} = \frac{1}{2}(v')^2 + \frac{1}{2}\frac{a''}{a}v^2 - \frac{1}{2}h^{ab}(\partial_a v)(\partial_b v)$$

and the Hamiltonian for conformal-time evolution is

$$H = \frac{1}{2}\int \mathrm{d}^3x\left(p_v^2 - \frac{a''}{a}v^2 + h^{ab}(\partial_a v)(\partial_b v)\right).$$

Also, here, we recognize the time-dependent frequency for the modes of v. Moreover, the canonical form allows us to quantize the modes individually and determine their fluctuations in a ground state. According to properties of the harmonic-oscillator ground state, with values of $m = 1$ for the mass and frequency $\omega = k$ at $\eta \to -\infty$ by the analogy used here, $\Delta v_k = \sqrt{\hbar/2k}$. The usual time dependence $e^{-i\omega t}$ of the asymptotic quantum state is then $e^{-ik\eta}$, associated with the parameter A in the general solution (4.129). We have an asymptotic vacuum state if we choose $B = 0$ and $A = \sqrt{\hbar/2k}$.

Transforming back to φ-perturbations, we obtain values of all the modes $\delta\varphi_k = v_k/a$ at later times, in particular at the time when a mode leaves the Hubble radius. From then on, its value will be frozen owing to the conservation of curvature perturbations, until it re-enters the Hubble radius long after inflation has ended. Values of the modes at Hubble exit thus provide the primordial spectrum of perturbations used for subsequent evolution. The distribution is usually expressed in terms of the power spectrum

$$P_{\delta\varphi}(k) = \frac{k^3}{2\pi^2}|\delta\varphi_k|^2_{\text{Hubble exit}} = \frac{k^3}{2\pi^2}|v_k/a|^2_{k=a\mathcal{H}}. \tag{4.130}$$

From (4.129) with $B = 0$, we see that $k^3|v_k/a|^2 = a^{-2}k^3A^2\left(1 + (k\eta)^{-2}\right) = \frac{1}{2}\hbar a^{-2}k^2\left(1 + a^2\mathcal{H}^2/k^2\right)$ from the de Sitter relationship $1/\eta = -a\mathcal{H}$ seen in Example 2.1. At horizon exit of the mode k, implying that $k = a\mathcal{H}$, we have $P_{\delta\varphi}(k) \propto \mathcal{H}^2$ which is nearly independent of k as long as $\mathcal{H}$ is almost constant, as assumed for the near-de Sitter derivation. The power spectrum is thus nearly scale-free.

Exercises

4.1 Consider a Lie algebra with generators X_I and structure constants C^I_{JK} such that $[X_J, X_K] = C^I_{JK} X_I$. Prove that the Jacobi identity

$$[X_I, [X_J, X_K]] + [X_K, [X_I, X_J]] + [X_J, [X_K, X_I]] = 0$$

implies that $C^L_{[IJ} C^M_{K]L} = 0$.

For structure constants

$$C^I_{JK} = \epsilon_{NJK} n^{(IN)} + \delta^I_K a_J - \delta^I_J a_K \,,$$

rewrite the Jacobi identity as $n^{IJ} a_J = 0$.

4.2 Let n^a be the unit timelike normal to a family of homogeneous spatial slices with spatial tangent spaces spanned by three linearly independent Killing vector fields ξ^a_I.

(i) Define the tensor $K_{ab} := \nabla_a n_b$ and show that all its non-spatial components vanish, i.e. $n^a n^b K_{ab} = n^a \xi^b_I K_{ab} = n^b \xi^a_I K_{ab} = 0$. It thus agrees with the extrinsic-curvature tensor of the homogeneous spatial slices and is symmetric.

(ii) Decompose K_{ab} into its trace-free part σ_{ab} (the shear tensor) and its trace θ (expansion):

$$K_{ab} = \sigma_{ab} + \frac{1}{3}\theta h_{ab} \,.$$

Show that

$$\mathcal{H} := \frac{1}{3}\theta = \frac{1}{6N}\frac{\mathrm{d}}{\mathrm{d}t}\log\det h$$

which agrees with the Hubble parameter if the spatial metric h_{ab} is isotropic. Time derivatives are taken with respect to a time-evolution vector field $t^a = N n^a$, i.e. we choose a vanishing shift vector (or one given by a Killing vector field) to preserve spatial homogeneity.

(iii) Express the momentum p^{ab} of h_{ab} in terms of σ_{ab} and θ and show that

$$\dot{\theta} = -8\pi G \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{p^{ab} h_{ab}}{\sqrt{\det h}}\right) .$$

(iv) Use the canonical equations of motion, in the presence of a matter Hamiltonian H_{matter}, to derive

$$\frac{\dot{\mathcal{H}}}{N} = -\mathcal{H}^2 - \frac{1}{3}\sigma^{ab}\sigma_{ab} - \frac{4\pi G}{3}(\rho_{\mathrm{E}} + 3P_{\mathrm{E}}) \,. \tag{E4.1}$$

Verify also that for $h_{bc} = a^2 \delta_{bc}$ the average pressure agrees with the pressure (2.27) defined for an isotropic metric.

4.3 Generalize the isotropic singularity theorem (Chapter 2.4) to anisotropic models by using equation (E4.1) to show, assuming the strong energy condition to be satisfied, that $\mathcal{H}$ must diverge after some finite amount of proper time if $\dot{\mathcal{H}} < 0$ initially. The average scale $a := (\det h)^{1/6}$ then becomes zero. Use the continuity equation $\dot{\rho} = -3\mathcal{H}(\rho + P)$, which still holds in this form in the anisotropic case, to show that ρ must diverge after a finite amount of proper time if the strong energy condition holds.

[Hint: show that $\mathrm{d}(\rho a^2)/\mathrm{d}t$ is positive under the above assumptions.]

4.4 (i) Find the minimum of the curvature potential for the Bianchi IX model and show that it is obtained for isotropic metrics.

(ii) Verify that the spatial Ricci scalar of the Bianchi IX model for an isotropic metric $h_{IJ} = h\delta_{IJ}$ agrees with the result ${}^{(3)}R = 6a^{-2}$ obtained from the direct isotropic calculation (Exercise 2.2 in Chapter 2).

In polar coordinates (ϑ, φ) and with $0 \leq \rho < 2\pi$, explicit invariant 1-forms for a Bianchi IX model are

$$
\begin{aligned}
\boldsymbol{\omega}^1 &= 2\sin\vartheta\cos\varphi \mathrm{d}\rho + [\sin 2\rho\cos\vartheta\cos\varphi - (\cos 2\rho - 1)\sin\varphi]\mathrm{d}\vartheta \\
&\quad + [-\sin 2\rho\sin\varphi - (\cos 2\rho - 1)\cos\vartheta\cos\varphi]\sin\vartheta\mathrm{d}\varphi \\
\boldsymbol{\omega}^2 &= 2\sin\vartheta\sin\varphi \mathrm{d}\rho + [\sin 2\rho\cos\vartheta\sin\varphi + (\cos 2\rho - 1)\cos\varphi]\mathrm{d}\vartheta \\
&\quad + [\sin 2\rho\cos\varphi - (\cos 2\rho - 1)\cos\vartheta\sin\varphi]\sin\vartheta\mathrm{d}\varphi \\
\boldsymbol{\omega}^3 &= 2\cos\vartheta\mathrm{d}\rho - \sin 2\rho\sin\vartheta\mathrm{d}\vartheta + (\cos 2\rho - 1)\sin^2\vartheta\mathrm{d}\varphi
\end{aligned}
$$

(see Exercise 4.7 below). Use this to compute the isotropic metric $\mathrm{d}s^2 = \sum_I \omega_a^I \omega_b^I \mathrm{d}x^a \mathrm{d}x^b$ and relate h to the scale factor a.

4.5 (i) Determine all non-vanishing structure constants for the Bianchi type V model.

(ii) Compute the spatial Ricci scalar for a diagonal type V metric which is isotropic, $h_{IJ} = h\delta_{IJ}$, using the general expression for ${}^{(3)}R$ in terms of arbitrary structure constants.

(iii) Assume that the momentum is also isotropic, $p^{IJ} = P\delta^{IJ}$, and relate the momentum p_h of h to P. (The momentum of h can be read off from $\dot{h}_{IJ}p^{IJ} = \dot{h}p_h$ using the fact that h_{IJ} and p^{IJ} are conjugate.)

(iv) Specialize the homogeneous Hamiltonian constraint to isotropic type V models and verify that it coincides with the Hamiltonian constraint for an isotropic model of negative spatial curvature ($k = -1$).

(v) Show that all boundary terms arising in the derivation of equations of motion for the Bianchi type V model vanish when the metric is specialized to isotropic form. The isotropic reduction of the type V model (which is of class B) therefore does have a valid Hamiltonian formulation corresponding to that of the isotropic model of negative spatial curvature.

4.6 Show that the volume of any diagonal Bianchi class A model decreases monotonically if it is decreasing initially.

[Hint: derive and use $\frac{\mathrm{d}}{\mathrm{d}t}(p_\alpha e^{-2\alpha}) \geq 0$.]

4.7 Derive the left-invariant 1-forms quoted in Exercise 4.4 from the Maurer–Cartan form $(\boldsymbol{\theta}_{\rm MC})_a = \omega^I_a T_I = s^{-1}(\mathrm{d}s)_a$. Parameterize group elements $s \in \mathrm{SO}(3)$ by Euler angles: $s(\rho, \vartheta, \varphi) = R_3(\rho)R_1(\vartheta)R_3(\varphi)$ with rotation matrices R_3 and R_1 around the z- and x-axes, respectively.

4.8 Show that the invariant connection $\mathbf{A}_{S/F}$ on $S/F = \mathrm{SO}(3)/\mathrm{U}(1)$ can be gauged to be of the form (4.55). To do so, use the embedding $\iota\colon S/F \to S$ described by $\rho = \varphi/2$ in terms of Euler angles and the Maurer–Cartan form from the preceding exercise. In order to change the gauge to the form (4.55), show and make use of

$$\begin{aligned} g\tau_1 g^{-1} &= -\sin\varphi n^i_\varphi \tau_i + \cos\varphi n^i_\vartheta \tau_i \\ g\tau_2 g^{-1} &= \cos\varphi n^i_\varphi \tau_i + \sin\varphi n^i_\vartheta \tau_i \\ g\tau_3 g^{-1} &= n^i_x \tau_i \end{aligned}$$

with $g := \exp(\vartheta n^i_\varphi \tau_i) \in \mathrm{SU}(2)$ and the unit vectors n^i_x, n^i_ϑ and n^i_φ along polar-coordinate lines.

4.9 Show that the action of the Euclidean group on $\mathbb{R}^3$ defined in Example 4.8 is indeed a group action.

Verify that inverse elements in the Euclidean group are given as in Example 4.13.

4.10 Show that a spherically symmetric densitized triad (4.56) is equivalent to a spherically symmetric spatial metric (4.59), and relate E^x and E^φ to L and S.

4.11 Compute the gauge transformations generated by the spherically symmetric diffeomorphism constraint (4.66), compare with Lie derivatives along a 1-dimensional vector field with component N^x, and conclude that p_S and L are densities, while S and p_L are scalar.

4.12 Show that a 1-form on a 1-dimensional space transforms like a scalar density under orientation-preserving coordinate changes.

Compute the reduced diffeomorphism constraint in spherically symmetric Ashtekar–Barbero variables and determine the density weights of all U(1)-gauge invariant components A_x, A_φ, E^x and E^φ.

4.13 Solve the canonical equations for a spherically symmetric metric such that spatial slices are flat ($\mathrm{d}s^2 = \mathrm{d}r^2 + r^2(\mathrm{d}\vartheta^2 + \sin^2\vartheta\,\mathrm{d}\phi^2)$ on slices $T = \mathrm{const}$) and the lapse function is $N(T, r) = 1$. (This choice of gauge results in the Painlevé–Gullstrand space-time.)

Show that this agrees with the Schwarzschild form in coordinates $(t, r, \vartheta, \varphi)$ where $T(t, r) = t + f(r)$, and use this to relate integration constants to the mass.

4.14 Show that the bivector (4.74) appearing in the action of 2-dimensional dilaton gravity defines a Poisson tensor.

[Hint: the vector $\frac{1}{2}\epsilon_{ijk}\mathcal{P}^{kl}$ might be useful for compactness of the calculations.]

4.15 Show that the action (4.75) is left invariant by the transformations (4.78) if and only if $\mathcal{P}^{ij}$ is a Poisson tensor.

4.16 Compute $\mathcal{L}_{\vec{\xi}}\, {}^0\varphi$ and $\mathcal{L}_{\vec{\xi}}\delta\varphi$ for a scalar field $\varphi(x) = {}^0\varphi + \delta\varphi(x)$ and compare with gauge transformations generated by the diffeomorphism constraint.

Repeat for a scalar density $p(x) = {}^0p + \delta p(x)$.

4.17 Compute the canonical form of gauge transformations for a scalar field φ, as well as the background and linear equations of motion in conformal time.

4.18 Show that the linear perturbation equations (2.39)–(2.42) together with the linear Klein–Gordon equation (2.43) form a consistent set.

5
Global and asymptotic properties

Hyperbolicity, as verified for general relativity in Chapter 3.4.2, guarantees the existence of local solutions in terms of initial data, but not the existence of global ones at all times. When evolving for long time intervals, singularities can develop in the solution and prevent it from being extendable further. We have already seen examples in homogeneous solutions of Bianchi models and the simpler isotropic solutions of FLRW models. In these cases, for matter satisfying the strong energy condition, there were always reasonable initial values which led to solutions with a diverging Hubble parameter and expansion rate at some time in the future or the past. The Hubble parameter corresponds to the expansion of the *family of timelike geodesics followed by comoving Eulerian observers*, a concept which presents a useful perspective in the context of singularities. When this expansion parameter diverges, the geodesic family no longer defines a smooth submanifold of space-time but develops a caustic or focal point where, in a homogeneous geometry, all geodesics intersect. For every spatial point on a non-singular slice, there is exactly one such geodesic intersecting it; when all these geodesics simultaneously intersect at the singular slice, it means that the whole space collapses into a single point after a finite amount of proper time. At this time, the initial value problem breaks down and space-time cannot be extended further.

In general inhomogeneous space-times, similar conclusions about geodesics can be derived. However, the route from a caustic of geodesic families, which could simply be a harmless focusing event of matter or radiation without major implications for the surrounding space-time, to a space-time singularity is not as direct. In this chapter we will start with a general discussion of timelike geodesic congruences and then proceed to the extra input required for singularity theorems that can be derived from geodesic properties. A complete understanding of the structure of singularities, as of space-time itself, requires global and asymptotic aspects, which make up the second part of this chapter. Concrete models for the formation of singularities by matter collapse will be discussed in the context of black holes. Several notions introduced in this chapter have a relation to the constructions made for canonical gravity; the general viewpoint here will be wider, however.

5.1 Geodesic congruences

A *geodesic congruence* is a family of geodesics, defining a surface in space-time which can be 2-, 3- or 4-dimensional depending on whether the geodesics are emanating transversally from a curve, a 2-dimensional surface or a spatial region as the cross-section with a spatial slice. The entire surface defined by the family may be mathematically singular, i.e. have self-intersections. At any point of each geodesic the tangent vector is defined as usual, extended to a tangent vector field ξ^a on the surface defined by the congruence. If the tangent vector field is everywhere timelike, which we then normalize to $\xi^a \xi_a = -1$, we have a timelike geodesic congruence. The null case with $\xi^a \xi_a = 0$ is also of interest.

5.1.1 Timelike geodesic congruences

On the geodesic congruence, we define the tensor field $B_{ab} := \nabla_b \xi_a$ such that $\xi^a B_{ab} = 0 = \xi^b B_{ab}$ using the normalization condition and the geodesic property of the tangent vector field. If ξ^a were the unit normal vector field to a spatial surface, B_{ab} would be its extrinsic-curvature tensor (3.49). In general, however, the tangent vector field ξ^a of a timelike geodesic congruence may not be hypersurface orthogonal, i.e. there may be no slicing into spatial surfaces such that ξ^a is normal to the slices at each cross-section with the congruence. Thus, unlike in the case of extrinsic curvature the Frobenius theorem used in Chapter 3.2.2.2 to show the symmetry of K_{ab} is not applicable and we may have a non-zero antisymmetric part $B_{[ab]} =: \omega_{ab}$, called the *rotation* of the geodesic congruence.

For further decomposition of B_{ab}, we introduce the symmetric tensor $h_{ab} = g_{ab} + \xi_a \xi_b$ (which would be the induced spatial metric if ξ^a were hypersurface orthogonal to a family of spatial slices). The symmetric part of B_{ab} can then be decomposed into its trace, the *expansion* $\theta := B^{ab} h_{ab} = B^{ab} g_{ab}$, and its symmetric trace-free part, the *shear* $\sigma_{ab} := B_{(ab)} - \frac{1}{3}\theta h_{ab}$. Thus, the whole field

$$B_{ab} = \frac{1}{3}\theta h_{ab} + \sigma_{ab} + \omega_{ab} \tag{5.1}$$

is uniquely decomposed into a scalar, a symmetric and an antisymmetric tensor.

The tensor B_{ab} changes along the geodesic family in a characteristic way. Its rate of change obeys the general equation

$$\begin{aligned} \xi^c \nabla_c B_{ab} &= \xi^c \nabla_c \nabla_b \xi_a = \nabla_b (\xi^c \nabla_c \xi_a) - (\nabla_b \xi^c)(\nabla_c \xi_a) + R_{cba}{}^d \xi^c \xi_d \\ &= -B^c{}_b B_{ac} + R_{cba}{}^d \xi^c \xi_d \,. \end{aligned} \tag{5.2}$$

Taking the trace, we obtain the rate of change of expansion, which satisfies the *Raychaudhuri equation*

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\theta = \xi^c \nabla_c \theta = -\frac{1}{3}\theta^2 - \sigma_{ab}\sigma^{ab} + \omega_{ab}\omega^{ab} - R_{cd}\xi^c \xi^d \tag{5.3}$$

where τ is proper time along the geodesics. For the timelike geodesic congruence of co-moving observers in an isotropic space-time, or for the evolution of large-scale modes in the separate-universe approach, this is the Raychaudhuri Eq. (2.22) for the Hubble rate $\mathcal{H} = \frac{1}{3}\theta$.

If the geodesic congruence is hypersurface orthogonal, i.e. $\omega_{ab} = 0$, and the strong energy condition is satisfied, i.e. $R_{ab}\xi^a\xi^b \geq 0$, having an initially negative expansion $\theta = \theta_0 < 0$ at one time τ_0 implies that θ diverges after a finite amount of proper time bounded by $\tau \leq 3/|\theta_0|$. This follows in exactly the same way as the isotropic singularity theorem for the Hubble rate of an isotropic universe. However, the diverging expansion of a geodesic congruence does not immediately imply a space-time singularity as it did in Chapter 2.4. It merely means that all geodesics in the congruence converge to a single point, but space-time could remain regular. (For a homogeneous space-time we used the unit-normal congruence. If it converges, this implies a singularity because space then collapses to a single point.)

5.1.1.1 Maximization of proper time

In order to use the focusing of geodesics for conclusions about space-time singularities, the focal point must somehow be related to an emerging boundary of space-time, or some other cataclysmic event. The main tool is to demonstrate that, under additional conditions, geodesics cannot be extended further beyond the focal point of their congruence. If this is the case, test masses which follow geodesics in the congruence can serve as probes to test space-time only for a limited range beyond which no space-time to be probed can exist. Space-time itself must develop a boundary when there are geodesics that stop at a certain point and cannot be extended further. If this is an insurmountable boundary reached in a finite amount of proper time by test particles or observers, it presents a *space-time singularity*.

To analyze extendability, it is useful to consider timelike curves of *maximum arc-length* (or proper time) from a given initial surface Σ to a point p. Such curves tell us how far we can at best go in space-time. To identify curves of maximal length, we take a 1-parameter family of curves all emanating from Σ and intersecting in p as in Fig. 5.1. This provides a congruence of timelike curves $\gamma_\sigma(t)$ labelled by $\sigma \in \mathbb{R}$ such that $\gamma_\sigma(0) \in \Sigma$ and $\gamma_\sigma(1) = p$ for all σ. By a variational principle, we now determine which conditions a curve of maximal length must satisfy.

Along the family of curves, we define the tangent vector fields

$$\xi^a := \left(\frac{\partial}{\partial t}\right)^a \quad \text{and} \quad \eta^a := \left(\frac{\partial}{\partial \sigma}\right)^a \tag{5.4}$$

such that $\eta^a|_{t=0}$ is tangent to Σ and $\eta^a|_{t=1} = 0$. Moreover, because these are coordinate vector fields, we have $[\xi, \eta]^a = 0$. (At this point we do not require normalization, which might not be compatible with commutation of the vector fields.)

First variation Each curve in the family has a length $\ell(\sigma) = \int_0^1 \sqrt{-\xi_a\xi^a}\mathrm{d}t$, and we have to find the curve labeled by σ_0 such that $\ell(\sigma_0)$ is maximal. At this value of σ in the family, the first variation must vanish and the second variation be negative. For the first variation

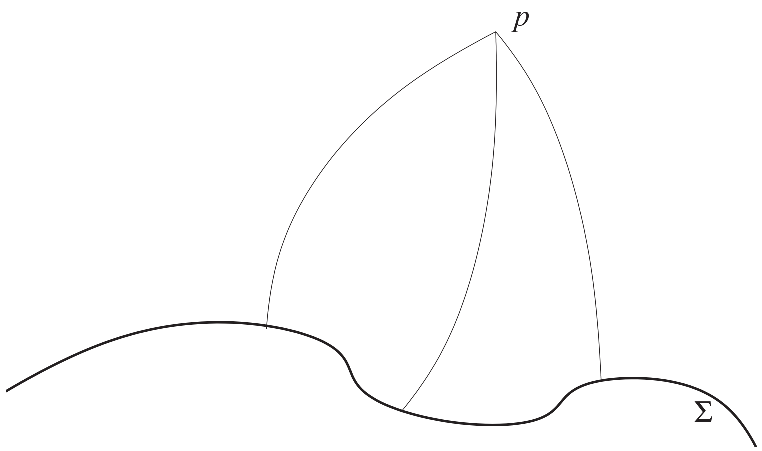

Fig. 5.1 Curves emanating from a spacelike surface Σ and meeting in a point p. By variation over the congruence of curves, the curve of maximum proper time can be determined.

by σ, using $\partial/\partial\sigma = \eta^a \nabla_a$, we have

$$\begin{aligned}
\frac{\mathrm{d}\ell}{\mathrm{d}\sigma} &= \int_0^1 \frac{-\xi_a \eta^b \nabla_b \xi^a}{\sqrt{-\xi_c \xi^c}} \mathrm{d}t = -\int_0^1 \frac{\xi_a \xi^b \nabla_b \eta^a}{\sqrt{-\xi_c \xi^c}} \mathrm{d}t \\
&= -\int_0^1 \xi^b \nabla_b \frac{\xi_a \eta^a}{\sqrt{-\xi_c \xi^c}} \mathrm{d}t + \int_0^1 \eta^a \xi^b \nabla_b \frac{\xi_a}{\sqrt{-\xi_c \xi^c}} \mathrm{d}t \\
&= -\left[\frac{\xi_a \eta^a}{\sqrt{-\xi_c \xi^c}} \right]_0^1 + \int_0^1 \eta^a \xi^b \nabla_b \frac{\xi_a}{\sqrt{-\xi_c \xi^c}} \mathrm{d}t \\
&= \left. \frac{\xi_a \eta^a}{\sqrt{-\xi_c \xi^c}} \right|_{t=0} + \int_0^1 \eta^a \xi^b \nabla_b \frac{\xi_a}{\sqrt{-\xi_c \xi^c}} \mathrm{d}t
\end{aligned} \tag{5.5}$$

where we used (5.4) and the fact that $\xi^a \nabla_a$ is a total t-derivative along the integration done here. The boundary term at $t = 1$ vanishes, since $\eta^a = 0$ there.

If a curve from Σ to p extremizes length, the variation of ℓ must vanish for all families of curves, i.e. for all vector fields η^a by which it can be deformed to neighboring curves in the family. Since η^a can be considered an arbitrary vector field tangent to the congruence, the boundary term as well as the integrand in (5.5), taken at all times, must vanish. The boundary term requires that $\xi_a \eta^a|_{t=0} = 0$, and the curve must emanate normally from Σ; the vanishing integral requires $\xi^b \nabla_b (\xi^a / \sqrt{-\xi_c \xi^c}) = 0$ along the extremizing curve. The acceleration of the unit tangent vector thus vanishes, exactly the condition for a geodesic. Extremal length is obtained for a geodesic normal to Σ.

Second variation To find out whether the extremum is a maximum or a minimum, we compute the second variation of ℓ around the extremizing curve, now known to be a geodesic normal to Σ. Starting from (5.5), the second variation is

$$\frac{\mathrm{d}^2\ell}{\mathrm{d}\sigma^2} = \int_0^1 \eta^c\nabla_c\left(\eta^a\xi^b\nabla_b\frac{\xi_a}{\sqrt{-\xi_d\xi^d}}\right)\mathrm{d}t + \eta^c\nabla_c\frac{\xi_a\eta^a}{\sqrt{-\xi_d\xi^d}}\bigg|_{t=0} =: I(\eta) + B(\eta)\,.$$

Note that it is meaningful to take the derivative of the boundary term along η^c because η^c is tangent to Σ at $t = 0$.

We will discuss the integral term $I(\eta)$ and the boundary term $B(\eta)$ separately. (The integral $I(\eta)$ is not zero in general because $\xi^b\nabla_b\frac{\xi_a}{\sqrt{-\xi_c\xi^c}}$ is required to vanish at one σ only, corresponding to the extremizing geodesic; its σ-derivative may be non-vanishing.)

For the integral term we have, using that γ_{σ_0} is the extremizing geodesic and that $[\xi,\eta]^a = 0$,

$$\begin{aligned}
I(\eta)|_{\sigma=\sigma_0} &= \int_0^1 \eta^c\eta^a(\nabla_c\xi^b)\nabla_b\frac{\xi_a}{\sqrt{-\xi_d\xi^d}}\mathrm{d}t + \int_0^1 \eta^c\eta^a\xi^b\nabla_c\nabla_b\frac{\xi_a}{\sqrt{-\xi_d\xi^d}}\mathrm{d}t \\
&= \int_0^1 \eta^a(\xi^c\nabla_c\eta^b)\nabla_b\frac{\xi_a}{\sqrt{-\xi_d\xi^d}}\mathrm{d}t + \int_0^1 \eta^c\eta^a\xi^b\nabla_b\nabla_c\frac{\xi_a}{\sqrt{-\xi_d\xi^d}}\mathrm{d}t \\
&\qquad + \int_0^1 \eta^c\eta^a\xi^b R_{cba}{}^d\frac{\xi_d}{\sqrt{-\xi_e\xi^e}}\mathrm{d}t \\
&= \int_0^1 \eta^a\xi^c\nabla_c\left(\eta^b\nabla_b\frac{\xi_a}{\sqrt{-\xi_d\xi^d}}\right)\mathrm{d}t + \int_0^1 \eta^c\eta^a\xi^b R_{cba}{}^d\frac{\xi_d}{\sqrt{-\xi_e\xi^e}}\mathrm{d}t \\
&= \int_0^1 \left(\eta^a\xi^c\nabla_c\left(\frac{\xi^b}{\sqrt{-\xi_d\xi^d}}\nabla_b\eta_a + \frac{\xi_a\xi^b\xi_d}{(-\xi_e\xi^e)^{3/2}}\nabla_b\eta^d\right)\right. \\
&\qquad \left. + \eta^c\eta^a\xi^b R_{cba}{}^d\frac{\xi_d}{\sqrt{-\xi_e\xi^e}}\right)\mathrm{d}t \\
&= \int_0^1 \eta^a\left(\xi^c\nabla_c\frac{\xi^b}{\sqrt{-\xi_d\xi^d}}\nabla_b\left(\eta_a + \xi_d\eta^d\frac{\xi^a}{-\xi_e\xi^e}\right)\right. \\
&\qquad \left. + R_{cba}{}^d\eta^c\xi^b\frac{\xi_d}{\sqrt{-\xi_e\xi^e}}\right)\mathrm{d}t \\
&= \int_0^1 \eta^{\perp a}\left(\xi^c\nabla_c\frac{\xi^b}{\sqrt{-\xi_d\xi^d}}\nabla_b\eta^\perp_a + R_{cba}{}^d\eta^{\perp c}\xi^b\frac{\xi_d}{\sqrt{-\xi_e\xi^e}}\right)\mathrm{d}t
\end{aligned} \tag{5.6}$$

having in the final steps again used the fact that ξ^a at the extremum is the tangent vector to a geodesic. We also defined the component

$$\eta^\perp_a = \eta_a - \frac{\eta^d\xi_d}{\xi_e\xi^e}\xi_a$$

of η_a orthogonal to ξ^a.

For the boundary term we have, using $\xi^a\eta_a = 0$ at $t = 0$,

$$
\begin{aligned}
B(\eta)|_{\sigma=\sigma_0} &= \eta^c\nabla_c \frac{\xi_a\eta^a}{\sqrt{-\xi_b\xi^b}}\bigg|_{t=0} \\
&= \frac{\xi_a}{\sqrt{-\xi_b\xi^b}}\eta^c\nabla_c\eta^a\bigg|_{t=0} + \frac{\xi_a\eta^a}{(-\xi_d\xi^d)^{3/2}}\xi_b\eta^c\nabla_c\xi^b\bigg|_{t=0} \\
&\quad + \frac{\eta^a}{\sqrt{-\xi_b\xi^b}}\eta^c\nabla_c\xi_a\bigg|_{t=0} \\
&= \frac{\xi_a}{\sqrt{-\xi_b\xi^b}}\eta^c\nabla_c\eta^a\bigg|_{t=0} + \eta^a\frac{\xi^c}{\sqrt{-\xi_b\xi^b}}\nabla_c\eta_a\bigg|_{t=0} \\
&= \frac{\xi_a}{\sqrt{-\xi_b\xi^b}}\eta^{\perp c}\nabla_c\eta^{\perp a}\bigg|_{t=0} + \eta^{\perp a}\frac{\xi^c}{\sqrt{-\xi_b\xi^b}}\nabla_c\eta^\perp_a\bigg|_{t=0} \\
&= -\eta^{\perp c}\eta^{\perp a}\nabla_c\frac{\xi_a}{\sqrt{-\xi_b\xi^b}}\bigg|_{t=0} + \eta^{\perp a}\frac{\xi^c}{\sqrt{-\xi_b\xi^b}}\nabla_c\eta^\perp_a\bigg|_{t=0} \\
&= -K_{ab}\eta^{\perp a}\eta^{\perp b}\big|_{t=0} + \eta^{\perp a}\frac{\mathrm{d}}{\mathrm{d}\tau}\eta^\perp_a\bigg|_{t=0}
\end{aligned}
\tag{5.7}
$$

with the extrinsic curvature K_{ab}, defined as in (3.49), of Σ as the surface at $t = 0$, whose unit normal is $\xi_a/\sqrt{-\xi_b\xi^b}$. Thus, $\eta^{\perp a}$ projects $\nabla_c(\xi_a/\sqrt{-\xi_b\xi^b})$ tangentially to Σ.

So far, we have assumed the curves to be differentiable. If there are finitely many kinks at points p_i where $\gamma_\sigma(t_i)$ is not differentiable and derivatives are discontinuous, we split the whole t-integration into a sum of integrals over smooth parts of curves. In particular, we will use the case of a kink in η^a while assuming that ξ^a remains smooth. Boundary terms of t-integrals from kink to kink then contribute via the second term in (5.7) the terms

$$
\Delta_i\left(\eta^{\perp a}\frac{\xi^b}{\sqrt{-\xi_c\xi^c}}\nabla_b\eta^\perp_a\right)
\tag{5.8}
$$

where we defined

$$
\Delta_i f := \lim_{\epsilon\to 0}(f(t_i+\epsilon) - f(t_i-\epsilon))
$$

for a function f. Thus, combining (5.6), (5.7) and (5.8) the second variation of arc length around the extremum given by a geodesic is

$$
\begin{aligned}
\frac{\mathrm{d}^2\ell}{\mathrm{d}\sigma^2}\bigg|_{\text{extremum}}(\eta) = \int_0^1 &\mathrm{d}\tau\,\eta^{\perp a}\left(\frac{\mathrm{d}^2}{\mathrm{d}\tau^2}\eta^\perp_a + R_{cba}{}^d\eta^{\perp c}\frac{\xi^b\xi_d}{-\xi_e\xi^e}\right) \\
&+ \eta^{\perp a}\frac{\mathrm{d}}{\mathrm{d}\tau}\eta^\perp_a\bigg|_{t=0} - K_{ab}\eta^{\perp a}\eta^{\perp b}\big|_{t=0} \\
&+ \sum_i \Delta_i\left(\eta^{\perp a}\frac{\mathrm{d}}{\mathrm{d}\tau}\eta^\perp_a\right).
\end{aligned}
\tag{5.9}
$$

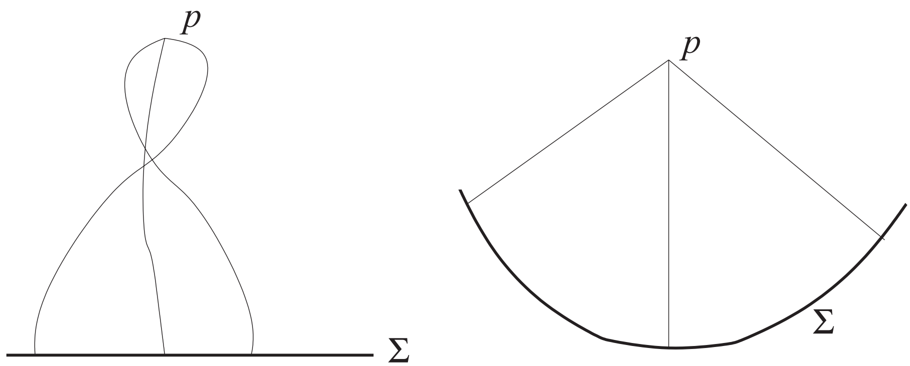

Fig. 5.2 Positive space-time curvature in the bulk or negative extrinsic curvature of the initial surface lead to focusing of geodesics which can result in focal points.

Maximization If we are allowed to ignore curvature and the boundary terms in (5.9), we have

$$\left.\frac{\mathrm{d}^2\ell}{\mathrm{d}\sigma^2}\right|_{\text{extremum}}(\eta) = \int_0^1 \mathrm{d}\tau\, \eta^{\perp a}\frac{\mathrm{d}^2}{\mathrm{d}\tau^2}\eta^{\perp}_a$$
$$= -\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}\tau}\left.\left(\eta^a\eta_a\right)\right|_{t=0} - \int_0^1\left(\frac{\mathrm{d}}{\mathrm{d}\tau}\eta^{\perp}_a\right)\left(\frac{\mathrm{d}}{\mathrm{d}\tau}\eta^{\perp a}\right)\mathrm{d}\tau\,. \qquad (5.10)$$

After integrating by parts, the boundary term at $t = 1$ vanishes, since $\eta^a = 0$ at p. The boundary term at $t = 0$ may be non-vanishing, but when it is small, the second variation cannot be positive: $\mathrm{d}\eta^{\perp}_a/\mathrm{d}\tau$ is normal to a timelike vector field ($\xi^a\mathrm{d}\eta^{\perp}_a/\mathrm{d}\tau = \mathrm{d}(\xi^a\eta^{\perp}_a)/\mathrm{d}\tau = 0$) and thus spacelike. We conclude that smooth timelike geodesics maximize arc-length in regions where curvature contributions can be ignored.

This presents a version of the twin paradox: the geodesic between two points always has the longest proper time, and any other curve which involves acceleration leads to slower aging. (In Euclidean signature, this corresponds to the well-known result that geodesics on a Riemannian manifold minimize arc-length. In Lorentzian signature, however, no curve minimizes arc-length because piecewise null curves produce arc-lengths arbitrarily close to zero.)

If the boundary term in (5.10) is sufficiently large, it may make the second variation positive. In this case, the curves emanating from the surface are strongly focused toward each other; the behavior at the boundary can thus compensate for some of the distance effects. Also, the other terms in (5.9) which we ignored so far are related to focusing effects as a consequence of curvature.

Globally, curvature effects must be considered and can change the maximization result. According to the two terms in the second variation (5.9), containing the space-time Riemann tensor and the extrinsic curvature tensor of the initial surface, there can be two different curvature effects contributing positive terms to $\mathrm{d}^2\ell/\mathrm{d}\sigma^2$; see Fig. 5.2: (i) positive space-time curvature in the form $R_{cba}{}^{d}\eta^{\perp c}\xi^{b}\eta^{\perp a}\xi_{d} > 0$ leads to bending and focusing of geodesics which then evolve into focal points, and (ii) negative extrinsic curvature of the initial surface would aim the geodesics toward each other, also implying a focal point even in the absence of space-time curvature. Beyond the focal point, timelike geodesics no longer maximize arc-length, which we now demonstrate using the notion of Jacobi vector fields.

5.1.1.2 Jacobi vector fields

A *Jacobi vector field* κ^{a} along a geodesic γ is a non-vanishing vector field which is everywhere orthogonal to the tangent vector ξ^{a}, i.e. $\xi^{a}\kappa_{a} = 0$, and is a solution of the *geodesic deviation equation*

$$\frac{\mathrm{d}^2}{\mathrm{d}\tau^2}\kappa_a + R_{cba}{}^{d}\kappa^{c}\frac{\xi^{b}\xi_{d}}{-\xi_{e}\xi^{e}} = 0\,. \tag{5.11}$$

This being a second-order differential equation for a 4-vector restricted by $\xi^{a}\chi_{a} = 0$, it provides six independent solutions for Jacobi vector fields. To recall, the geodesic deviation equation is satisfied by a deviation vector $(\partial/\partial\sigma)^{a}$ in a congruence $\gamma_{\sigma}(t)$ whose curves for fixed σ are all geodesic. One can thus interpret the family of Jacobi vector fields as a parameterization of all possible embeddings of the given geodesic γ in a congruence, an interpretation that illustrates our interest in Jacobi vector fields for the questions considered here.

Geodesic congruences For a congruence of geodesics emanating from a surface and culminating in a focal point, there must be a Jacobi vector field satisfying $\kappa^{a}(1) = 0$ if the focal point is reached at $t = 1$, as well as

$$\left.\frac{\mathrm{d}\kappa_a}{\mathrm{d}\tau}\right|_{t=0} = K_{ab}\kappa^{b}|_{t=0} \tag{5.12}$$

at the initial surface $\Sigma : t = 0$. We show this in the following calculation.

Example 5.1 (Geodesic congruences and extrinsic curvature)
For the proper-time derivative of a Jacobi vector field along a geodesic congruence, we have

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\kappa_a = \frac{\xi^{b}}{\sqrt{-\xi^{c}\xi_{c}}}\nabla_{b}\kappa_{a} = \frac{\kappa^{b}\nabla_{b}\xi_{a}}{\sqrt{-\xi^{c}\xi_{c}}}\,.$$

We did not normalize the tangent vector ξ^{a} to keep our coordinate vector fields (5.4) commuting, a property used in the second step of the preceding equation.

At the initial surface $\Sigma : t = 0$ of the congruence, the normalized tangent vector to geodesics provides the unit normal, since the deviation vector χ^{a} must be tangent to the surface and is normal to the geodesic tangent vectors in the case of a Jacobi vector field. The vector field ξ^{a}, once normalized, can thus be used to compute the extrinsic curvature

$$K_{ab} = h_{a}^{c}\nabla_{c}\frac{\xi_{b}}{\sqrt{-\xi^{d}\xi_{d}}}$$

of Σ. We compare the two formulas using also

$$\kappa^b \nabla_b \sqrt{-\xi^c \xi_c} = -\frac{\xi^c \kappa^b \nabla_b \xi_c}{\sqrt{-\xi^d \xi_d}} = -\frac{\xi^c \xi^b \nabla_b \kappa_c}{\sqrt{-\xi^d \xi_d}} = \frac{\kappa_c \xi^b \nabla_b \xi_c}{\sqrt{-\xi^d \xi_d}} = 0$$

at $t = 0$ because κ^a and ξ^a are orthogonal and since ξ^a is tangent to a geodesic. Thus,

$$\frac{\mathrm{d}\kappa_a}{\mathrm{d}\tau} = \kappa^b \nabla_b \frac{\xi_a}{\sqrt{-\xi^c \xi_c}} = \kappa^b K_{ab} \,,$$

proving (5.12), since κ^a is tangent to Σ at $t = 0$, providing a projection of the derivative of the unit tangent.

If a Jacobi vector field κ^a with the boundary conditions $\kappa^a(1) = 0$ and (5.12) exists along the geodesic γ from Σ to $q = \gamma(1)$, the point $q = \gamma(1)$ is called *conjugate* to Σ.

We assume now that we have a geodesic γ from Σ to a point p, and that it lies in a congruence with a conjugate point q to Σ before p is reached along γ. Thus, between Σ and q we have a Jacobi vector field with the above boundary conditions; in particular, $\kappa^a = 0$ at q. We continuously extend κ^a as a vector field beyond q by saying that it vanishes between q and p. Moreover, we choose another vector field λ^a with $\xi^a \lambda_a = 0$ such that $\lambda_a \mathrm{d}\kappa^a/\mathrm{d}\tau = -1$ at q. Using $\eta^a := \epsilon\lambda^a + \epsilon^{-1}\kappa^a$ for some $\epsilon \in \mathbb{R}\backslash\{0\}$, we define the deviation vector field of a new congruence (not necessarily geodesic, and one with a kink at q where η^a is not smooth). It implies a second variation of arc-length

$$\begin{aligned} \left.\frac{\mathrm{d}^2\ell}{\mathrm{d}s^2}\right|_{\text{extremum}} (\eta) &= \epsilon^2 \left.\frac{\mathrm{d}^2\ell}{\mathrm{d}s^2}\right|_{\text{extremum}} (\lambda) + 2\Delta_q(\lambda^a \mathrm{d}\kappa_a/\mathrm{d}\tau) \\ &= \epsilon^2 \left.\frac{\mathrm{d}^2\ell}{\mathrm{d}s^2}\right|_{\text{extremum}} (\lambda) + 2 \end{aligned}$$

using the fact that κ^a is piecewise Jacobi (or identically zero beyond q) with the given boundary condition (5.12) at Σ, and taking one new contribution in (5.9) from the point q where the congruence has its kink. This variation becomes positive for sufficiently small ϵ, which proves that the timelike geodesic γ no longer provides a curve of maximum length beyond a focal point (see Fig. 5.3).

Diverging expansion Jacobi fields can also be used to demonstrate that the expansion of a congruence diverges at a focal point. We pick three independent Jacobi fields $\kappa_i^a, i = 1, 2, 3$, satisfying the boundary condition (5.12) at Σ. They give us the deviation vectors of three independent congruences. The components of κ_i^a can be viewed as forming a matrix, which must be invertible at places of the congruence that are not conjugate points to Σ. Then, from $(\xi^b/\sqrt{-\xi^c\xi_c})\nabla_b\kappa_i^a = \kappa_i^b\nabla_b(\xi^a/\sqrt{-\xi^c\xi_c}) = \kappa_i^b B_b{}^a$ (by a calculation as in Example 5.1) we obtain the expansion

$$\theta = B^a{}_a = (\kappa^{-1})^i_a \frac{\xi^b}{\sqrt{-\xi^c\xi_c}} \nabla_b \kappa_i^a = \frac{\xi^b}{\sqrt{-\xi^c\xi_c}} \nabla_b \log|\det(\kappa_i^a)| \,. \tag{5.13}$$

At a conjugate point, the matrix κ_i^a becomes degenerate, and the expansion diverges.

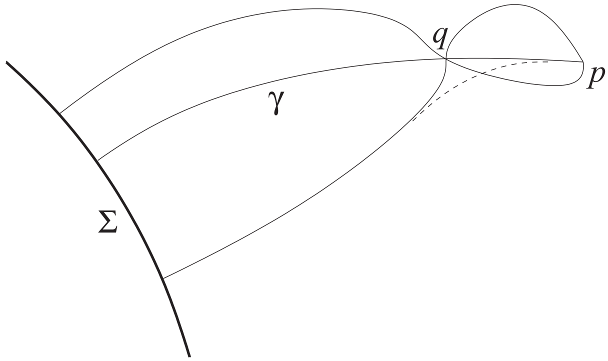

Fig. 5.3 Beyond a focal point, non-geodesic curves of proper time larger than that of any geodesic exist (dashed). (The Euclidean representation here makes the dashed extension look shorter, not longer.)

Just as in the case of isotropic and homogeneous singularity theorems, the Raychaudhuri equation guarantees the existence of conjugate points in congruences emanating from a surface whose extrinsic curvature has a trace bounded from above by a negative constant, $K^a{}_a \leq C < 0$, provided that the strong energy condition holds. The tangent vector field ξ^a on the congruence is normal to the initial surface, where we have $K^a{}_a = \nabla^a(\xi_a/\sqrt{-\xi^b\xi_b}) = \theta \leq C < 0$. Moreover, ξ^a is hypersurface orthogonal at Σ by assumption, and thus has initially vanishing rotation. It then follows from the rate of change (5.2) of B_{ab} that $\omega_{ab} = 0$ on the whole congruence as long as θ is regular, i.e. at least until a conjugate point is reached. Thus, the Raychaudhuri equation (5.3) can be estimated and solved just as in the homogeneous case in Chapter 2.4, with the conclusion that θ diverges after a proper time of at most $\tau \leq 3/|C|$; there must be a conjugate point before that time.

We are now able to formulate the first general singularity theorem (for details, see Hawking and Ellis (1973)):

Theorem 5.1 (Singularity theorem) *Assume that the strong energy condition is satisfied and that space-time is globally hyperbolic. If there is a spacelike Cauchy surface Σ of extrinsic curvature $K^a{}_a \leq C < 0$, then no timelike curve from Σ can have a proper length larger than $3/|C|$.*

Proof (Sketch) The proof is indirect: assume that there is a timelike curve λ with proper length larger than $3/|C|$ and take a point p on λ of a distance more than $3/|C|$ away from Σ. Global hyperbolicity can be shown to imply the existence of a curve γ of maximum length from Σ to p, which by assumption must have a length larger than $3/|C|$. From the previous results, this can only be a geodesic without a conjugate point between Σ and p, in contradiction to what follows from the Raychaudhuri equation. □

The theorem demonstrates that under the given conditions space-time is incomplete. There are timelike curves of finite proper time that cannot be extended further. Test masses following these curves only exist for a finite amount of proper time.

5.1.2 Null geodesic congruences

Null congruences, whose tangent vector field is everywhere null, require a decomposition of B_{ab} different from (5.1). If we again define $\tilde{h}_{ab} = g_{ab} + k_a k_b$, now calling the null tangent vector field k^a, it is degenerate on vectors orthogonal to k^a, including k^a itself. Formally, one can still define the decomposition as in (5.1), since no inverse of h_{ab} is required. However, this will not capture the correct number of degrees of freedom describing a null congruence.

5.1.2.1 Decomposition

There are two equivalent ways to deal with the degeneracy. For the first one, as followed by Wald (1984), we consider the 2-dimensional factor space of all vectors orthogonal to k^a modulo the vector space K generated by k^a: any element in this factor space can be represented by $\kappa^a + \lambda k^a$ where $\kappa^a k_a = 0$ and $\lambda \in \mathbb{R}$ is arbitrary. The degenerate direction is here simply factored out. Any tensor $T^{a_1\ldots a_n}_{b_1\ldots b_m}$ that vanishes on K, i.e. when contracted with k^a on any of its indices, yields a well-defined projection to the factor space, then denoted by a hat: $\hat{T}^{a_1\ldots a_n}_{b_1\ldots b_m}$. Due to $T^{a_1\ldots a_n}_{b_1\ldots b_m} k_{a_i} = 0 = T^{a_1\ldots a_n}_{b_1\ldots b_m} k^{b_j}$ for all $1 \leq i \leq n$ and $1 \leq j \leq m$ by assumption, a contraction of the tensor with $\kappa^a + \lambda k^a$ on any one of its indices will not depend on λ, and thus will yield the same result for the whole equivalence class of vectors. The projected tensor $\hat{T}^{a_1\ldots a_n}_{b_1\ldots b_m}$ is in this way defined on the factor space. One example is $\hat{h}_{ab}$ obtained from $\tilde{h}_{ab}$ (or g_{ab}, since the projection of $k_a k_b$ vanishes) as the non-degenerate induced metric on the 2-dimensional factor space.

For the second procedure, we note that the null property of the tangent vector field implies the existence of a second independent null vector field ℓ^a, $\ell^a \ell_a = 0$, which we can cross-normalize with k^a such that $\ell^a k_a = -1$. Then, $h_{ab} := g_{ab} + 2\ell_{(a} k_{b)}$ satisfies $\ell^a h_{ab} = 0 = k^a h_{ab}$ and is thus the 2-metric on a 2-dimensional subspace normal to both k^a and ℓ^a. Defining $\hat{h}_{ab}$ as the evaluation of h_{ab} on the 2-dimensional subspace of the tangent space normal to both k^a and ℓ^a, we obtain a specific representative of the factor-space metric defined by the first procedure.

Another example for a projectable tensor is $B_{ab} = \nabla_b k_a$: it satisfies $B_{ab} k^a = 0 = B_{ab} k^b$ as already seen in the timelike case. (Here, we have to use the affine parameterization for null geodesics, i.e. $k^c \nabla_c k^a = 0$.) The projection is then decomposed as

$$\hat{B}_{ab} = \frac{1}{2} \theta \hat{h}_{ab} + \hat{\sigma}_{ab} + \hat{\omega}_{ab} \tag{5.14}$$

with

$$\theta = g^{ab} \hat{B}_{ab} \,, \qquad \hat{\sigma}_{ab} = \hat{B}_{ab} - \frac{1}{2} \theta \hat{h}_{ab} \,, \qquad \hat{\omega}_{ab} = \hat{B}_{[ab]} \,. \tag{5.15}$$

(Notice that we have the trace $g^{ab} \hat{h}_{ab} = g^{ab}(g_{ab} + 2\ell_a k_b) = \delta^a_a + 2\ell^a k_a = 2$.) Since the affine parameterization is fixed only up to a multiplicative constant of k^a, also θ, $\hat{\sigma}_{ab}$ and $\hat{\omega}_{ab}$ are defined only up to a multiplicative constant in the null case.

As before, we derive

$$k^c \nabla_c \hat{B}_{ab} + \hat{B}^c{}_b \hat{B}_{ac} = \widehat{R_{cbad} k^c k^d}$$

which implies that

$$k^c \nabla_c \theta = -\frac{1}{2}\theta^2 - \hat{\sigma}_{ab}\hat{\sigma}^{ab} + \hat{\omega}_{ab}\hat{\omega}^{ab} - R_{cd} k^c k^d \,. \tag{5.16}$$

The last term is again non-positive if the strong energy condition holds, but now even for the weak energy condition, since k^c is null. We conclude that there must be a focal point in any hypersurface-orthogonal null congruence whose expansion θ_0 is negative at some value of the affine parameter, and this focal point is reached within an affine length of at most $2/|\theta_0|$. (This statement is invariant under rescaling the affine parameter λ defined by $k^c \nabla_c = \mathrm{d}/\mathrm{d}\lambda$: λ and $1/\theta_0$ scale in the same way if the affine parameter is changed. Suitable combinations such as $\lambda\theta_0$ are scaling invariant.)

5.1.2.2 Trapped surfaces and singularities

As in the timelike case, a singularity theorem requires us to connect the appearance of a focal point in geodesic congruences with inextendability. While timelike congruences capture the intuitive notion of co-moving observers experiencing the whole expansion or collapse of space in a cosmological situation, null congruences describe the motion of light. For cosmological singularities, having a time when the expansion is everywhere bounded away from zero was seen to guarantee a singularity in the past or the future. Null congruences are related to a different concept that provides an initial condition to guarantee singularities. Focal points of null geodesic congruences lead to singularities if a *trapped surface* T exists: a compact, 2-dimensional smooth spacelike submanifold T such that the expansion of *both* sets of future-directed null geodesics orthogonal to T is *everywhere negative* on T.

In Minkowski space, any compact 2-dimensional spacelike surface has one expanding ("outgoing") and one contracting ("ingoing") family of future-directed normal null congruences. But even light rays shining "out of" a trapped surface are contracted toward smaller cross-sections of the beam. A geometry allowing trapped surfaces must thus be rather different from the one known for Minkowski space. After showing in the rest of this subsection the importance of trapped surfaces for singularities, recognized by Penrose (1965), we will provide examples and interpretations for them in what follows.

Theorem 5.2 (Singularity theorem: trapped-surface version) *Assume that the weak or strong energy condition holds and that space-time M is globally hyperbolic with a non-compact connected Cauchy surface Σ. If a trapped surface T exists with expansions $\theta \leq \theta_0 < 0$ for both normal null congruences, then there is an inextendible null geodesic of affine distance no larger than $2/|\theta_0|$ from T.*

Proof (Sketch) We first introduce the future $I^+(T)$ of T as the set of all points in M connected to T by future-directed timelike curves. Its boundary $\dot{I}^+(T)$ is closed (the boundary of a boundary is empty) and is a 3-dimensional (but not necessarily differentiable;

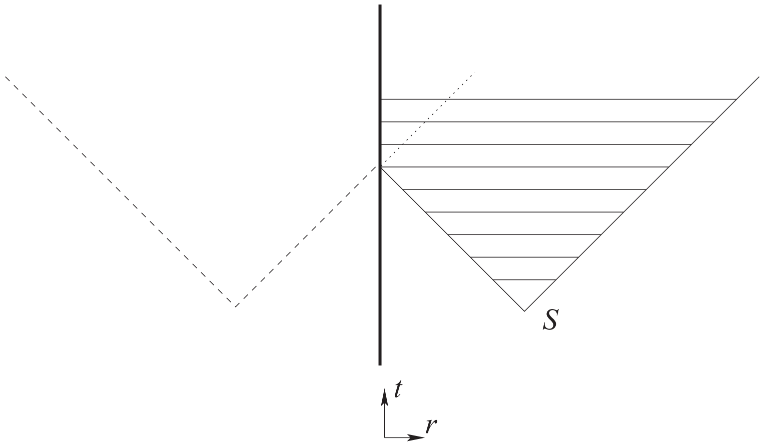

Fig. 5.4 The future (hashed) of a spherical surface has a non-smooth part where it intersects the center. Here, we show a 2-dimensional space-time diagram in which points represent 2-surfaces in space-time and light rays travel along $\pm 45^0$ lines. (More on this technique of causal diagrams will be discussed in Chapter 5.2.2.) To the future of the kink, light signals appear to originate from the reflection of the initial sphere S at the $r = 0$ world-line (dashed). The first reflection point is a conjugate point of S, and lies at a kink of $\dot{I}^+(S)$. Except for this point, $r = 0$ is not part of $\dot{I}^+(S)$; the future of the kink is in the interior of $I^+(S)$. The boundary of the future is non-differentiable: points in the diagram are spheres for $r \neq 0$ but points also in 4-dimensional space-time for $r = 0$. Including the dashed mirror-reflection in the diagram illustrates the non-smooth behavior of $\dot{I}^+(S)$ in space-time.

see Fig. 5.4) manifold locally generated by null geodesics. In general, $\dot{I}^+(T)$ does not contain complete null geodesics: any point on a null geodesic beyond a conjugate point can be connected to T by a timelike curve and thus lies in the interior rather than the boundary of $I^+(T)$. (This can be shown by constructing a deformation similar to that used in Chapter 5.1.1.2 to show that timelike geodesics fail to maximize on arc-length beyond conjugate points.)

In the context of the theorem, the expansion of null geodesics emanating from a trapped surface T is initially bounded from above by a negative value. Using the Raychaudhuri equation, all null geodesics have a conjugate point after some finite affine length. Provided that all null geodesics are extendible arbitrarily, $\dot{I}^+(T)$ must then be compact. (It is closed as the boundary of the future and bounded because the affine distance between any two points in the set is bounded. Note that it does not have a boundary, but like a 2-sphere in space is bounded in the sense that its complement in space-time is open.) To bring this to a contradiction, we use an arbitrary time-evolution vector field whose trajectories define a map $\psi\colon \dot{I}^+(T) \to S \subset \Sigma$ which is a homeomorphism; see Fig. 5.5. The image S of a compact set must then be closed and bounded but cannot have a boundary, as these topological properties are transferred from $\dot{I}^+(T)$: no homeomorphism can map a compact to a non-compact set. This contradicts the non-compactness of the Cauchy surface Σ assumed in

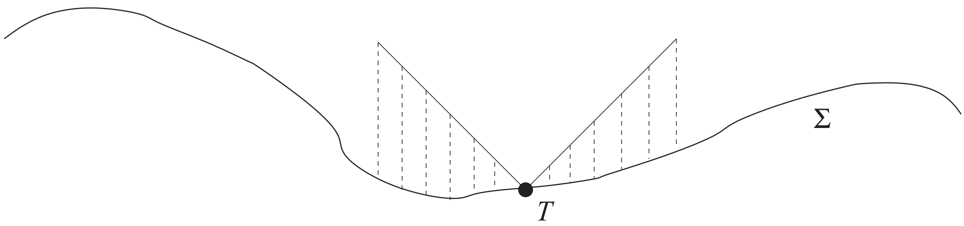

Fig. 5.5 Mapping the future of T to a spatial surface Σ along trajectories of a time-evolution vector field.

the theorem. The only consistent conclusion is that our assumption of extendability of all null geodesics must be violated, and we again derive geodesic incompleteness. □

Many assumptions in the singularity theorems can be weakened if more refined techniques are used. For instance, global hyperbolicity is not necessary. Four essential conditions then remain, as reviewed, e.g., by Senovilla (1998):

- an energy condition;
- a genericity assumption (such as the one stating that each timelike geodesic contains a point where $R_{abcd}\xi^a\xi^d \neq 0$);
- there is no closed timelike curve; and
- at least one additional property specifying an initial condition:
  - (i) a closed universe (a compact spatial submanifold without boundary exists),
  - (ii) a trapped surface exists, or
  - (iii) there is a point p such that the expansion of future-directed null geodesics from p becomes negative along each geodesic.

Under these conditions, space-time must be singular in the sense that at least one inextendible timelike or null geodesic exists. (In general, however, it is not guaranteed that all such geodesics are inextendible, nor are conclusions about curvature divergences drawn.)

At such a general level, the actual conclusions of singularity theorems may appear rather weak, predicting properties much tamer than what is known from explicit examples in cosmology or the physics of black holes. Moreover, the conclusions can be evaded by violating some of the more peculiar assumptions even without introducing exotic matter forms, as first constructed by Senovilla (1990); see also the detailed description by Chinea et al. (1992). Extensions of the traditional assumptions and theorems have thus been attempted, for instance by Clarke (1998), Vickers and Wilson (2000, 2001) and Senovilla (2007), but the level of generality of the theorems based on geodesic incompleteness remains unsurpassed. Nevertheless, singularity theorems are of utmost conceptual importance: they show that there is no mechanism in classical general relativity that could guarantee unlimited evolution of generic initial values. The theorems highlight major roadblocks to complete evolution within the theory, and they show the stability of singularities: they refer only to inequalities in their assumptions, such that perturbations around explicitly known singularities cannot eliminate singular behavior.

5.2 Trapped surfaces

We have seen two versions of singularity theorems, one of which is suitable for cosmological solutions, and one for black holes:

1. Extrinsic curvature on a Cauchy slice which is bounded from above by a negative constant means that the universe is globally contracting at one time. The singularity Theorem 5.1 implies geodesic incompleteness which one can interpret as a consequence of the universe collapsing into a singularity.
2. The existence of a trapped surface means that there is a bounded region of space-time where the gravitational field is strong enough to contain even light emanating from the trapped surface. (In this case light is contained at least locally, as the negative value of expansion of the outgoing null normal congruence on the trapped surface may grow and become positive away from the surface.) According to singularity Theorem 5.2, this region collapses into a singularity which one can associate with a black hole.

In both cases, the fact that gravity is purely attractive under most circumstances (unless pressure becomes very negative) removes any means to stop the collapse of matter and space once it has progressed sufficiently far.

5.2.1 Black-hole solutions

We have seen the cosmological situation of collapse in several examples of homogeneous models. The trapped-surface version of the singularity theorem is realized in the simplest case for the Schwarzschild solution

$$\mathrm{d}s^2 = -\left(1-\frac{2M}{r}\right)\mathrm{d}t^2 + \left(1-\frac{2M}{r}\right)^{-1}\mathrm{d}r^2 + r^2\mathrm{d}\Omega^2 \,. \tag{5.17}$$

This line element is ill-defined at $r = 2M$ owing to a failure of the coordinate system used. The two separate regions $r > 2M$ and $r < 2M$ can, however, be discussed appropriately in Schwarzschild coordinates. Their behavior is quite different:

$r > 2M$: A timelike Killing vector field $\partial/\partial t$ exists in addition to the spherical Killing vector fields. The metric in this region is stationary (and, in fact, static).

$r < 2M$: Because $1 - 2M/r < 0$ in this region, the Killing vector field $\partial/\partial t$ is now spacelike. The metric is no longer stationary but rather dynamical. Instead of stationarity, the Killing vector field together with spherical symmetry implies homogeneity: there are three independent spacelike Killing vectors at each point.[1]

[1] The submanifolds defined by $r = \text{const}$ are spacelike and have the topology $\mathbb{R} \times S^2$ with t as the coordinate of $\mathbb{R}$ and the polar coordinates on the 2-spheres. Since there are three independent rotational Killing vector fields (only two of which are linearly independent at each point), in addition to homogeneity, there is a rotational symmetry around one axis transverse to the 2-spheres. The geometry is of Kantowski–Sachs type, a homogeneous model not contained as a special case in the Bianchi classification: its symmetry group is 4-dimensional without allowing a 3-dimensional Bianchi subgroup.

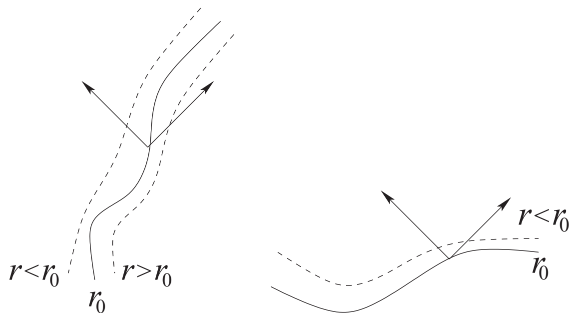

Fig. 5.6 If $r = r_0 =$ const defines timelike surfaces, the two future-pointing null normal congruences to spheres at r_0 either increase or decrease the cross-section area; the spheres are thus untrapped (left). If $r = r_0 =$ const is spacelike, however, both normal congruences change cross-section areas in the same way — the region is trapped (right).

In particular, surfaces $r =$ const are timelike for $r > 2M$ but spacelike for $r < 2M$. This property has immediate consequences for the existence of trapped surfaces, as visualized in the causal diagram Fig. 5.6 where null lines run at $\pm 45^0$: in any situation in which r is the area radius coordinate such that the angular part of the metric is $r^2 \mathrm{d}\Omega^2$ as in (5.17), a *spacelike* surface $r =$ const implies that 2-dimensional spherical cross-sections defined by $r =$ const, $t =$ const are *trapped*. If $r =$ const is timelike, future-pointing null normals reach smaller r for the "ingoing" family and larger r for the "outgoing" family. The first congruence has negative expansion and the second one positive expansion because the cross-section area proportional to r^2 increases for the latter. A surface at constant t is thus not trapped. But if $r =$ const is spacelike, both future pointing null normals can only reach either smaller or larger r. It is then only possible that both expansions are negative or both are positive. Spheres at constant t and r are trapped (future trapped if smaller r is to the future, and past trapped if larger r is to the future).

Example 5.2 (Spherical trapped surfaces)

In order to see more concretely that the nature of $r =$ const *surfaces in a spherically symmetric space-time provides a criterion for spherical trapped surfaces, we compute the general form of radial null vector fields in a 2-dimensional space-time with general line element* $\mathrm{d}s^2 = g_{tt}\mathrm{d}t^2 + 2g_{rt}\mathrm{d}t\mathrm{d}r + g_{rr}\mathrm{d}r^2$. *We assume that* $g_{rt} > 0$ *and that* t *is a good time coordinate in the whole region considered; if there is a flip of* t *and* r *as happens when the horizon in Schwarzschild coordinates is traversed, the following arguments must be adapted. A radial null vector* ξ^a *satisfies* $g_{rr}(\xi^r)^2 + 2g_{rt}\xi^r\xi^t + g_{tt}(\xi^t)^2 = 0$, *which we interpret as a quadratic polynomial for* ξ^r. *Null vector fields thus satisfy one of the two relationships*

$$\xi^r_\pm = -\frac{g_{rt}}{g_{rr}}\left(1 \mp \sqrt{1 - \frac{g_{tt}}{g_{rr}g_{rt}^2}}\right)\xi^t_\pm \qquad (5.18)$$

between their components. For a future-pointing null vector, we assume that $\xi^t_\pm > 0$*, but for the following argument we do not need to restrict* $\xi^t_\pm$ *further. (For a null geodesic family in affine parameterization, for instance, the geodesic equation would provide another condition for* $\xi^t_\pm(t,r)$ *to satisfy.)*

A null normal congruence emanating from a sphere at constant $r = r_0$ *is expanding if its tangent vector field* ξ^a *satisfies* $\xi^a \partial_a r > 0$*, and is contracting if* $\xi^a \partial_a r < 0$*. By this condition, the vector field* ξ^a_- *in (5.18) is always contracting while* ξ^a_+ *may be expanding or contracting depending on the geometry. It is expanding if* $g_{tt}/g_{rr} < 0$*, and contracting if* $g_{tt}/g_{rr} > 0$*. The intermediate case belongs to a spherical marginally trapped surface. Inverting the line element, the condition can equally be formulated in terms of* $g^{rr}/g^{tt} = g_{tt}/g_{rr}$ *whose sign in regions where the lapse function does not vanish, and thus* $g^{tt} = -1/N^2 < 0$*, is solely determined by* $g^{rr} = g^{ab}(\mathrm{d}r)_a(\mathrm{d}r)_b = |n_a|^2$ *with the co-normal* $n_a = (\mathrm{d}r)_a$ *to spheres. Thus, if the normal* n^a *to* $r = r_0$ *is spacelike (such that the surface* $r = r_0$ *is timelike),* $g^{rr} > 0$ *and the outgoing null normal congruence is expanding. The surface* $r = r_0$ *is not trapped. In the opposite case, for a timelike normal* n^a *both normal null congruences are contracting and* $r = r_0$ *is trapped.*

Horizon properties from different perspectives especially in the spherical context are exhibited by Nielsen and Visser (2006) and Nielsen (2009).

Example 5.3 (Trapped surfaces in Painlevé–Gullstrand coordinates)
Both regions of the Schwarzschild space-time can be described simultaneously in a coordinate system of Painlevé–Gullstrand form where (5.17) becomes:

$$\mathrm{d}s^2 = -\mathrm{d}T^2 + \left(\mathrm{d}r + \sqrt{2M/r}\,\mathrm{d}T\right)^2 + r^2\mathrm{d}\Omega^2 \tag{5.19}$$

with a suitable new coordinate $T(t,r)$*. (As an aside, notice that spatial slices of constant* T *are flat.) Also here, for* $r < 2M$ *surfaces* $r =$ const *are spacelike because they have the induced metric* $\mathrm{d}s^2 = -(1-2M/r)\mathrm{d}T^2 + r^2\mathrm{d}\Omega^2$*. We can explicitly compute null congruences with tangents* $\xi^a_\pm$ *normal to spheres from the conditions* $g_{ab}\xi^a_\pm\xi^b_\pm = 0$ *together with* $\xi^a_\pm\nabla_a T > 0$ *to pick the future-pointing ones. This yields*

$$\xi^a_\pm = \left(\frac{\partial}{\partial T}\right)^a - \left(\sqrt{\frac{2M}{r}} \mp 1\right)\left(\frac{\partial}{\partial r}\right)^a .$$

With $\xi^a_-\nabla_a r = -(1+\sqrt{2M/r}) < 0$*,* ξ^a_- *is always the ingoing null normal of negative expansion. For the outgoing null normal* ξ^a_+*, we have* $\xi^a_+\nabla_a r = 1 - \sqrt{2M/r}$ *which is negative for* $r < 2M$*. Thus, in this region both future-pointing null normal congruences have negative expansion because the area radius decreases along both of them. Here, spheres of constant* r *and* T *are trapped.*

For the sherically symmetric Schwarzschild solution, one can of course construct the whole space-time solution and directly verify the existence of a curvature singularity to the future of the trapped region $r < 2M$: r plays the role of a time coordinate for $r < 2M$

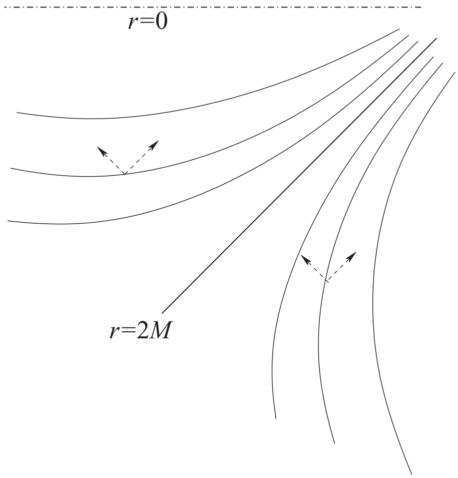

Fig. 5.7 Surfaces of constant r in the Schwarzschild regions. At $r = 2M$, the behavior of the $r = \text{const}$ surfaces changes from timelike outside to spacelike inside. The region for $r < 2M$ is trapped and is bounded by a singularity at $r = 0$ to the future (dash-dotted). How the singularity is related to the null line $r = 2M$ and other (possibly infinite) boundaries of space-time will be determined by conformal completions below.

and progresses to the future singularity at $r = 0$ as seen in Fig. 5.7. The key advantage of a singularity theorem is that it demonstrates the stability of this conclusion, at least as far as singularities due to geodesic incompleteness are concerned. The singularity is not just a consequence of a highly symmetric spherical collapse but is more general: even if one perturbs around the Schwarzschild solution, specific values of expansions of null congruences may change, but for sufficiently small perturbations they will remain negative if they are negative for the unperturbed solution. This demonstrates that a whole class of non-symmetric solutions has trapped surfaces and must be geodesically incomplete as a consequence of Theorem 5.2.

5.2.2 Causal diagrams

In spherical symmetry, properties of conjugate points and the relation to trapped surfaces can be visualized in causal pictures. We ignore angular coordinates and thus draw a 2-dimensional diagram of space-time where time increases vertically. The symmetry axis at $r = 0$ is drawn as a vertical boundary, at which incoming light rays are reflected: coming from positive r, they reach the center $r = 0$ and return to positive r. While this is drawn as a reflection in the 2-dimensional diagram, it simply means that the light ray traverses the center without any major physical event (Fig. 5.8).

A point at the boundary must be conjugate to any sphere at positive r because here spherical null congruences emanating from spheres surrounding the center converge to a point as seen in Fig. 5.8. As a Jacobi vector field, we can use the deviation vector of a radial

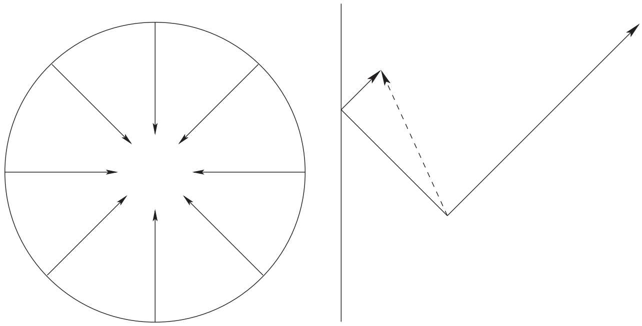

Fig. 5.8 The center as a conjugate point of a sphere in a spatial (left) and causal view (right). Beyond the conjugate point, a timelike curve can be used to connect to the original surface (dashed).

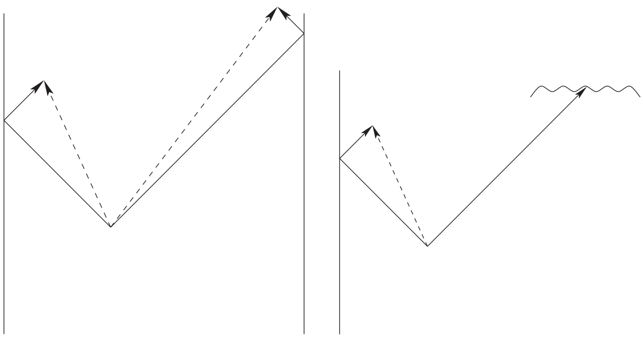

Fig. 5.9 A trapped surface in a compact space-time (left), where conjugate points are reached in any direction, and in a non-compact one (right), where the trapped surface implies a singularity.

congruence; indeed, it vanishes at the center. Extending a null ray beyond the conjugate point, there are points on the original sphere closer to the endpoint of the light ray than the point where it started. It is thus possible to connect the endpoint to the sphere by a timelike curve (moving at a speed less than that of light), in accordance with our general statement about conjugate points of null geodesic congruences made earlier (Fig. 5.8, right).

For a trapped surface, both the ingoing and outgoing null normal congruences have conjugate points if their geodesics are arbitrarily extendable. In a globally hyperbolic and connected space-time, this is only possible if space is compact, as illustrated in Fig. 5.9. In this case, moving around space along a null curve implies that one must reach a point which can be connected to the starting point by a timelike curve, followed by a particle avoiding the reflection and thus not having to travel as fast as light on the null ray. If space is not compact, however, it is not possible for both null normal congruences to have conjugate points. If a trapped surface exists, a contradiction can be avoided only by concluding that not all null geodesics, in particular those in congruences not reaching a conjugate point, can be extended. This is the intuitive content of the trapped-surface singularity Theorem 5.2.

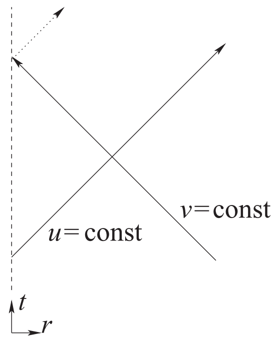

Fig. 5.10 Double-null coordinates u and v in Minkowski space and their coordinate lines. At the center (dashed), a $v = \text{const}$ line is "reflected" into a $u = \text{const}$ line, and vice versa.

5.3 Asymptotic infinity

For a full view on space-time, the behavior at infinity is also important. There is a rich structure, owing to the fact that the behavior depends on the direction in which one approaches infinity, for instance along null curves compared to spacelike curves. Structures at infinity also provide useful tools to extract physical and observable information: light from non-trapped regions can leave strong-curvature regimes and provide signals to distant observers. How one can extract observables from geodesic congruences or space-time geometries is an important problem for relativistic astrophysics. The situation here is much more complicated than in early-universe cosmology, where simple gauge-invariant combinations of linear metric and matter perturbations provide the main observable quantities.

To start with describing structures at infinity, in particular the space-time metric near infinity, the infinite boundary can first be mapped to a finite one via a coordinate transformation. For example, for Minkowski space with line element $\mathrm{d}s^2 = -\mathrm{d}t^2 + \mathrm{d}r^2 + r^2\mathrm{d}\Omega^2$ we may introduce the null coordinates $v := t + r$ and $u := t - r$ restricted to the range $v \geq u$ because $r \geq 0$ (see Fig. 5.10). In these coordinates, the metric is

$$\mathrm{d}s^2 = -\mathrm{d}u\mathrm{d}v + \frac{1}{4}(v-u)^2\mathrm{d}\Omega^2 \,. \tag{5.20}$$

We are interested in the limits $v \to \infty$, $u = \text{const}$ (outgoing null rays) and $u \to -\infty$, $v = \text{const}$ (incoming null rays). These are the only null lines due to $v \geq u$; $v \to -\infty$ at constant u and $u \to \infty$ at constant v are not possible.

The asymptotic limiting points are mapped to finite values by using a new set of coordinates, $T = \arctan v + \arctan u$, $R = \arctan v - \arctan u$, such that $\mathrm{d}T = \mathrm{d}v/(1+v^2) + \mathrm{d}u/(1+u^2)$ and $\mathrm{d}R = \mathrm{d}v/(1+v^2) - \mathrm{d}u/(1+u^2)$, or

$$-\mathrm{d}T^2 + \mathrm{d}R^2 = -\frac{4}{(1+u^2)(1+v^2)}\mathrm{d}u\mathrm{d}v \,.$$

This is not the radial part of Minkowski space (5.20), but it suggests a conformal transformation to $\tilde{g}_{ab} = \Omega^2 g_{ab}$ with

$$\Omega^2 = \frac{4}{(1+u^2)(1+v^2)} \,. \tag{5.21}$$

As the new metric conformal to Minkowski space, we thus obtain

$$\mathrm{d}\tilde{s}^2 = \Omega^2 \mathrm{d}s^2 = -\mathrm{d}T^2 + \mathrm{d}R^2 + \sin^2 R \mathrm{d}\Omega^2 \tag{5.22}$$

where we used

$$\begin{aligned}\sin R &= \sin(\arctan v)\cos(\arctan u) - \cos(\arctan v)\sin(\arctan u)\\ &= \frac{1}{\sqrt{(1+v^{-2})(1+u^2)}} - \frac{1}{\sqrt{(1+v^2)(1+u^{-2})}} = \frac{v-u}{\sqrt{(1+u^2)(1+v^2)}} \,.\end{aligned}$$

The metric (5.22) is that of an FLRW model with positive spatial curvature and a constant scale factor. Such a solution, called *Einstein static universe*, exists in the presence of a cosmological constant; see Exercise 2.3 in Chapter 2. The construction shows that the whole Minkowski space is conformally equivalent to a *finite* region of the Einstein static universe with metric (5.22), where T and R are bounded. Infinity of Minkowski space can be studied by a neighborhood of the boundary of the region it defines in the Einstein static universe. This finite region is characterized by the conditions (i) $R \geq 0$ due to $v \geq u$ and (ii) $-\pi < T \pm R < \pi$ due to properties of the arctan used in the definitions of T and R. Moreover, $R < \pi$. When the finite region representing the original space-time is combined with its boundary in a conformal embedding, the closed region thus obtained is called *conformal completion* or a *Carter–Penrose diagram*.

The boundary of the region in the Einstein static universe conformally equivalent to Minkowski space is of triangular shape (not smooth but with three corners). One can distinguish several characteristic components; see Fig. 5.11:

$R = 0, -\pi \leq T \leq \pi$: this is a timelike line representing the center of polar coordinates. It has two endpoints, one called i^+ at $T = \pi$ and one called i^- at $T = -\pi$.
$R = \pi, T = 0$: a single point called i^0.
$T + R = \pi$: a null curve at constant v, called $\mathcal{J}^+$ (and pronounced "scri-plus").
$T - R = -\pi$: a null curve at constant u, called $\mathcal{J}^-$.

The meaning of these boundary regions can be seen from the behavior of radial geodesics approaching them. In Minkowski space, these are straight lines represented by linear relationships between t and r:

Timelike geodesics Radial, future-pointing timelike geodesics satisfy $t = ar + b$ with some $|a| > 1$ near $t \to \infty$. Thus, $t \pm r = (1 \pm a^{-1})t \mp b/a \to \infty$, and both u and v approach infinity at the future endpoint in Minkowski space. In the conformally embedded region, this leads to $T \to \pi$ and $R \to 0$, which is i^+, or *future timelike infinity*.

Similarly, past-pointing timelike geodesics have the above form for $t \to -\infty$, where u and v approach $-\infty$. Here, $T \to -\pi$ and $R \to 0$ approaches i^-, or *past timelike infinity*.

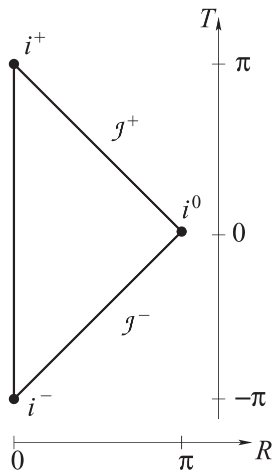

Fig. 5.11 Conformal completion of Minkowski space with its boundary components, embedded in the Einstein Static Universe.

Radially outgoing spacelike geodesics Here, we have the form $t = ar + b$ with $|a| < 1$ near $r \to \infty$. Thus, $t \pm r = (a \pm 1)r + b \to \pm\infty$ and $v \to \infty$, $u \to -\infty$. In this way, we reach $T \to 0$, $R \to \pi$ and thus i^0, or *spacelike infinity*.

Radially outgoing null geodesics The form $t = r + b$ near $t \to \infty$ implies $v \to \infty$ while $u = b$ remains constant. Thus, $T + R \to \pi$, $|T - R| < \pi$ and we reach $\mathscr{J}^+$, or *future null infinity*.

Radially incoming null geodesics In this case, $t = -r + b$ near $t \to -\infty$. Now, $v = b$ remains constant while $u \to -\infty$, implying $T - R \to -\pi$, $|T + R| < \pi$. All these geodesics come from $\mathscr{J}^-$, or *past null infinity*.

Cataclysmic events such as black holes lead to conformal completions very different from the global structure of Minkowski space, impressively showing the difference in geometries.

5.3.1 Asymptotic flatness

Although we have explicitly constructed the conformally completed space-time region only for Minkowski space, the form of asymptotic infinity obtained is general for a large class of space-times, or at least certain parts of them. Let us take an arbitrary space-time (M, g_{ab}) and assume that it can be conformally embedded in $(\tilde{M}, \tilde{g}_{ab} = \Omega^2 g_{ab})$ as a finite region whose boundary is given by the set where $\Omega = 0$, in such a way that $\tilde{\nabla}_a \Omega \neq 0$ is non-vanishing and finite wherever the boundary is differentiable. The original space-time is said to be *asymptotically empty* if its energy momentum tensor behaves as $T_{ab} = O(\Omega^3)$ near $\Omega = 0$. These and other asymptotic definitions, together with techniques for the construction of general conformal diagrams, go back to Penrose (1963).

In order to show that all such space-times have the same structure at infinity as Minkowski space, we will use a more general version to allow for the presence of a cosmological

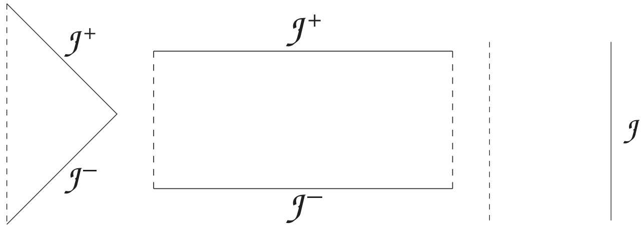

Fig. 5.12 Conformal diagrams of an asymptotically flat, asymptotically de Sitter and asymptotically anti de Sitter space. Dashed lines indicate symmetry centers or periodic identifications.

constant. In this case, T_{ab} will not approach zero near the boundary and the asymptotic structure will change, but we can use the same calculations to determine its properties. More generally, we assume that there is a constant Λ such that $T_{ab} + \Lambda g_{ab}/8\pi G = O(\Omega^3)$ near $\Omega = 0$. Einstein's equation, satisfied on M, then implies a Ricci scalar of the form

$$\begin{aligned} R &= -8\pi G g^{ab} T_{ab} = -8\pi G \Omega^{-2} \tilde{g}^{ab} T_{ab} \\ &= 4\Lambda + O(\Omega) \end{aligned} \tag{5.23}$$

where we used the fact that the components of $\tilde{g}^{ab}$ are finite at $\Omega = 0$. Moreover, the conformal transformation provides a relation between the Ricci scalars of g_{ab} and $\tilde{g}_{ab}$, derived directly from the expression of R in terms of the metric based on equations such as (3.118):

$$\tilde{R} = \Omega^{-2} \left(R - 6 g^{ab} \nabla_a \nabla_b \log \Omega - 6 g^{ab} (\nabla_a \log \Omega)(\nabla_b \log \Omega) \right) \tag{5.24}$$

or, inversely,

$$R = \Omega^2 \tilde{R} + 6\Omega \tilde{g}^{ab} \tilde{\nabla}_a \tilde{\nabla}_b \Omega - 12 \tilde{g}^{ab} (\tilde{\nabla}_a \Omega)(\tilde{\nabla}_b \Omega)\,. \tag{5.25}$$

Because all individual factors on the right-hand side remain finite for $\Omega \to 0$, only the last term is non-zero at $\Omega = 0$. We see that

$$\mathrm{sgn}\left((\tilde{\nabla}_a \Omega)(\tilde{\nabla}^a \Omega) \right)|_{\Omega=0} = -\mathrm{sgn} R|_{\Omega=0} = -\mathrm{sgn}\Lambda\,. \tag{5.26}$$

The gradient $\tilde{\nabla}_a \Omega = n_a$ is the co-normal to the boundary $\Omega = 0$, and so the boundary is

- null if $\Lambda = 0$, in which case the space-time (M, g_{ab}) is called *asymptotically Minkowski*,
- spacelike if $\Lambda > 0$, in which case the space-time (M, g_{ab}) is called *asymptotically de Sitter*, and
- timelike if $\Lambda < 0$, in which case the space-time (M, g_{ab}) is called *asymptotically anti de Sitter*.

The conformal diagrams look like the examples in Fig. 5.12.

To eliminate possible pathologies, one often defines geometries more specific than those corresponding asymptotically to any one of the maximally symmetric ones.

Definition 5.1 *A space-time (M, g_{ab}) is* asymptotically simple *if a subset $N \subset M$ and a space-time $(\tilde{M}, \tilde{g}_{ab})$ exist such that*

(i) *M is an open submanifold of $\tilde{M}$;*
(ii) *there is a real function Ω on $\tilde{M}$ such that $g_{ab} = \Omega^2 \tilde{g}_{ab}$ on M and $\Omega = 0$, $\tilde{\nabla}_a \Omega \neq 0$ on the smooth part of the boundary of M in $\tilde{M}$;*
(iii) *every null geodesic in $M \backslash N$ has an endpoint on the boundary of M in $\tilde{M}$.*

Moreover, (M, g_{ab}) is asymptotically flat *if it is asymptotically simple and $R_{ab} = 0$ near the boundary in $\tilde{M}$ (i.e. it is asymptotically simple and asymptotically empty).*

The second condition implies that the set defined by $\Omega = 0$ in $\tilde{M}$ is at infinity as seen from M. Choose a null geodesic $\tilde{\gamma}$ in $\tilde{M}$ that is affinely parameterized such that $\tilde{\lambda} = 0$ at $\Omega = 0$. Near the boundary, the affine parameter can be chosen to behave like $\tilde{\lambda} \sim \Omega$, since $\tilde{\nabla}_a \Omega \neq 0$. The equation

$$\frac{\mathrm{d}^2 x^a}{\mathrm{d}\tilde{\lambda}^2} + \tilde{\Gamma}^a_{bc} \frac{\mathrm{d}x^b}{\mathrm{d}\tilde{\lambda}} \frac{\mathrm{d}x^c}{\mathrm{d}\tilde{\lambda}} = 0$$

for the geodesic in $\tilde{M}$ implies that

$$\frac{\mathrm{d}^2 x^a}{\mathrm{d}\lambda^2} + \Gamma^a_{bc} \frac{\mathrm{d}x^b}{\mathrm{d}\lambda} \frac{\mathrm{d}x^c}{\mathrm{d}\lambda} = -\frac{1}{\mathrm{d}\lambda/\mathrm{d}\tilde{\lambda}} \left(\frac{\mathrm{d}^2\lambda}{\mathrm{d}\tilde{\lambda}^2} \left(\frac{\mathrm{d}\lambda}{\mathrm{d}\tilde{\lambda}} \right)^{-1} + \frac{2}{\Omega} \frac{\mathrm{d}\Omega}{\mathrm{d}\tilde{\lambda}} \right) \frac{\mathrm{d}x^a}{\mathrm{d}\lambda}$$

where we used the conformal transformation rule

$$\tilde{\Gamma}^a_{bc} = \Gamma^a_{bc} + 2\Omega^{-1} \delta^a_{(b} \nabla_{c)} \Omega - \Omega^{-1} (\nabla_d \Omega) g^{ad} g_{bc} \tag{5.27}$$

(analogous to (3.118)) as well as the null condition $g_{bc}(\mathrm{d}x^b/\mathrm{d}\lambda)(\mathrm{d}x^c/\mathrm{d}\lambda) = 0$. The affine parameter λ for the geodesic in M thus satisfies

$$\frac{\mathrm{d}^2\lambda}{\mathrm{d}\tilde{\lambda}^2} \left(\frac{\mathrm{d}\lambda}{\mathrm{d}\tilde{\lambda}} \right)^{-1} + \frac{2}{\Omega} \frac{\mathrm{d}\Omega}{\mathrm{d}\tilde{\lambda}} = 0$$

or $\mathrm{d}\lambda/\mathrm{d}\tilde{\lambda} = c\Omega^{-2}$. Since $\tilde{\lambda} \sim \Omega$ near $\Omega = 0$, $\lambda \sim c \int \Omega^{-2} \mathrm{d}\Omega$ diverges at $\Omega = 0$. Null geodesics in M never reach the boundary in a finite affine parameter distance; the boundary appears to be at infinity as seen from within M.

The third condition in Definition 5.1 eliminates possibly trapped and singular regions. The necessity of subtracting a subset N is due to spatially closed null geodesics never reaching infinity, as they may exist even in non-trapped regions of a general space-time. In the Schwarzschild space-time, for instance, null curves at constant r and $\theta = \pi/2$ have a tangent k^a satisfying $-(1 - 2M/r)(k^t)^2 + r^2 \sin^2\theta (k^\phi)^2 = 0$ as a relation between the t and ϕ-components of a spatially closed null curve along a ϕ-coordinate line. One can choose the components of the tangent to depend only on r and θ and thus be constant along the null curve. As for the geodesic equations, $k^a \nabla_a k^t = 0 = k^a \nabla_a k^\phi$ are identically satisfied, and $k^a\nabla_a k^\theta = -\sin\theta\cos\theta(k^\phi)^2$ vanishes at $\theta = \pi/2$. Finally,

$$\begin{aligned} k^a\nabla_a k^r &= \Gamma^r_{tt}(k^t)^2 + \Gamma^r_{\phi\phi}(k^\phi)^2 \\ &= \frac{M}{r^2}\left(1-\frac{2M}{r}\right)(k^t)^2 - r\left(1-\frac{2M}{r}\right)\sin^2\theta(k^\phi)^2 \\ &= -(r-3M)r^2\sin^2\theta(k^\phi)^2 \end{aligned}$$

vanishes for $r = 3M$ as the radius of a spatially closed null geodesic (indeed lying outside the trapped region $r < 2M$).

5.3.2 Examples

In general, finding suitable transformations to conformally embed a given space-time as a finite region of some other space-time can be challenging. Should this be the case for a space-time of interest, one can make use of a procedure patching together the whole space-time out of parts that can each be conformally embedded. Alternatively, the behavior of null-lines can be used to determine the structure of a conformal diagram, as applied by Winitzki (2005). The general construction is illustrated by the usual examples of explicitly known solutions.

5.3.2.1 Schwarzschild exterior

With

$$u := t - \left(r + 2M\log\left(\frac{r}{2M}-1\right)\right) \quad \text{and} \quad \ell := \frac{1}{r} \tag{5.28}$$

the Schwarzschild metric becomes

$$\mathrm{d}s^2 = \frac{1}{\ell^2}\left(-\ell^2(1-2M\ell)\mathrm{d}u^2 + 2\mathrm{d}u\mathrm{d}\ell + \mathrm{d}\Omega^2\right).$$

(An analogously defined v instead of u, adding instead of subtracting the r-dependent term, would bring the metric to Eddington–Finkelstein form when expressed in coordinates (v, r, θ, ϕ).) In the exterior, we have $0 < \ell < 1/2M$, and $\ell \to 0$ at $r \to \infty$. A conformal factor $\Omega = \ell$ thus satisfies the required conditions for a conformal completion, including $\nabla_a\Omega \neq 0$ at $\Omega = 0$. Moreover, since $R_{ab} = 0$ near $\ell = 0$, we have an asymptotically empty space-time as shown in Fig. 5.13.

In contrast to Minkowski space, the past of $\mathcal{J}^+$ constructed in this way is not the whole space-time. Instead, the past has a boundary given by the *event horizon* at $r = 2M$. (By definition, the event horizon in a given space-time is the boundary of the past of $\mathcal{J}^+$.) Furthermore, not all timelike geodesics reach i^+ in the future; those crossing the horizon enter the trapped region and stop at the singularity, the place of their inextendability.

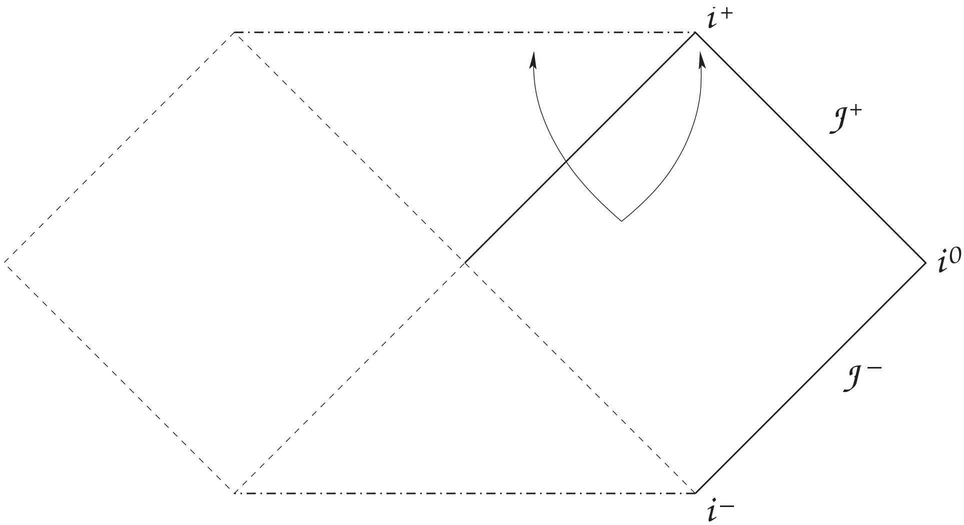

Fig. 5.13 Conformal diagram of the Schwarzschild space-time. In Eddington–Finkelstein coordinates, only the region spanned by the solid lines and the future singularity is covered. The future fate of objects differs from that in Minkowski space: depending on the initial radial velocity, a massive object may fall into the singularity or reach future timelike infinity. Similarly, a massless particle traveling along null curves either escapes to $\mathcal{J}^+$ or falls into the singularity.

5.3.2.2 Spatially closed FLRW model

In conformal time η, an FLRW line element with spatial slices of positive curvature is

$$\mathrm{d}s^2 = a(\eta)^2 \left(-\mathrm{d}\eta^2 + \mathrm{d}R^2 + \sin^2 R \mathrm{d}\Omega^2\right) . \tag{5.29}$$

This is already conformally equivalent in an obvious way to the Einstein static universe. In the absence of a cosmological constant, η has a bounded range because there is a past as well as a future singularity at finite η where $a(\eta)$ vanishes. The radial coordinate R is not restricted by singularities, but is periodically identified due to the spatially closed topology. The conformal diagram is shown in Fig. 5.14.

A recollapsing, spatially closed model is not asymptotically flat as a consequence of the homogeneous matter distribution, never diluting down arbitrarily much. For this reason, it does not have the structure of null, timelike and spacelike infinities.

5.3.2.3 Open FLRW model

An open model in conformal time does not directly provide a conformal embedding into the Einstein static universe, but rather, in Minkowski or hyperbolic space. Qualitatively, in the case of a vanishing cosmological constant, we now expect a regular part of the boundary given by future null and timelike infinity, the asymptotic region where matter has diluted completely, but also a past singularity. These ingredients, together with the coordinate center as a timelike finite boundary, already provide the conformal diagram in Fig. 5.14.

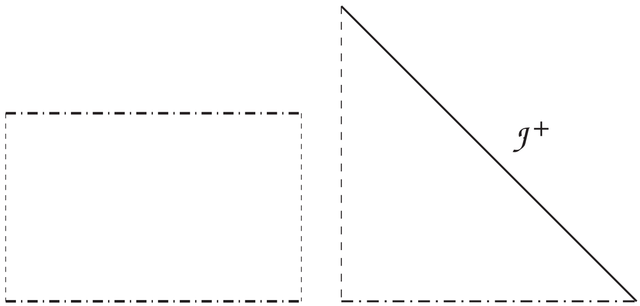

Fig. 5.14 Conformal diagrams of closed (left) and open (right) FLRW models.

5.3.3 Conformal gauge

For a conformal completion, one must choose a conformal embedding such that the space-time investigated becomes a bounded region in the conformally equivalent space-time. The choice of conformal embedding is not unique: every $\tilde{g}'_{ab} = \omega^2 \tilde{g}_{ab}$ with $\omega(x) \neq 0$ on $\tilde{M}$ provides the same asymptotic structure for M. This freedom is called *conformal gauge*.

For the structures at infinity, the normal to the boundary of M in $\tilde{M}$ has played an important role by telling us whether boundary components are null, spacelike or timelike. At $\mathcal{J}^\pm$, changing the conformal gauge by a function ω as above transforms the co-normal $n_a = \nabla_a \Omega$ on the boundary to $n'_a = \omega \nabla_a \Omega + \Omega \nabla_a \omega = \omega n_a$ (since $\Omega = 0$ at the boundary), and $\tilde{\nabla}_a n_b$ becomes

$$\begin{aligned}
\tilde{\nabla}'_a n'_b &= \tilde{\nabla}_a n'_b - 2\omega^{-1} \delta^c_{(a} (\tilde{\nabla}_{b)} \omega) n'_c + \omega^{-1} (\tilde{\nabla}_d \omega) \tilde{g}^{cd} \tilde{g}_{ab} n'_c \\
&= \omega \tilde{\nabla}_a n_b + n_b \tilde{\nabla}_a \omega + (\tilde{\nabla}_a \Omega)(\tilde{\nabla}_b \omega) + \Omega \tilde{\nabla}_a \tilde{\nabla}_b \omega - 2 n_{(a} \tilde{\nabla}_{b)} \omega + \tilde{g}_{ab} n^c \tilde{\nabla}_c \omega \\
&= \omega \tilde{\nabla}_a n_b + \tilde{g}_{ab} n^c \tilde{\nabla}_c \omega
\end{aligned}$$

where the last equality holds on $\mathcal{J}^\pm$. (In the first line, we have conformally transformed the connection using (5.27).)

While there was no restriction on $\tilde{\nabla}_a n_b$ so far, this transformation of the conformal gauge can be used to set $\tilde{\nabla}'_a n'_b$ to equal any particular function on the boundary, for instance making the simple choice $\tilde{\nabla}'_a n'_b = 0$ called the Bondi gauge. By this extra condition, the conformal gauge will be restricted, requiring ω to satisfy the equation $\tilde{g}_{ab} n^c \tilde{\nabla}_c \omega = -\omega \tilde{\nabla}_a n_b$.

In the case of null infinity, in which n^a is both normal and tangent to the boundary, this equation becomes a differential equation along $\mathcal{J}^\pm$. To rewrite $\tilde{\nabla}_a n_b$ as it appears in the differential equation, we make use of (5.25) and

$$R_{ab} = \tilde{R}_{ab} + 2\Omega^{-1} \tilde{\nabla}_a \tilde{\nabla}_b \Omega + \tilde{g}_{ab} \tilde{g}^{cd} \left(\Omega^{-1} \tilde{\nabla}_c \tilde{\nabla}_d \Omega - 3\Omega^{-2} (\tilde{\nabla}_c \Omega)(\tilde{\nabla}_d \Omega) \right) . \tag{5.30}$$

Given a sufficiently strong fall-off of T_{ab} and using Einstein's equation, (5.25) implies that

$$\Omega^{-1} \tilde{g}^{cd} (\tilde{\nabla}_c \Omega)(\tilde{\nabla}_d \Omega) = \frac{1}{2} \tilde{g}^{cd} \tilde{\nabla}_c \tilde{\nabla}_d \Omega$$

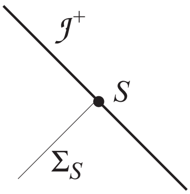

Fig. 5.15 Future null infinity and a cut.

on $\mathcal{J}^{\pm}$. (The first term in (5.25), $\Omega^2\tilde{R}$ with regular $\tilde{R}$, shows that $6\Omega\tilde{g}^{ab}\tilde{\nabla}_a\tilde{\nabla}_b\Omega - 12\tilde{g}^{ab}(\tilde{\nabla}_a\Omega)(\tilde{\nabla}_b\Omega)$ vanishes on $\mathcal{J}^{\pm}$ and remains finite when divided by Ω.) The Ricci tensor relation $R_{ab} = 0$ near $\mathcal{J}^{\pm}$ on M then implies that

$$\frac{1}{4}\tilde{g}_{ab}\tilde{g}^{cd}\tilde{\nabla}_c\tilde{\nabla}_d\Omega = \tilde{\nabla}_a n_b\,.$$

Thus, we can write

$$\tilde{\nabla}'_a n'_b = \omega\tilde{\nabla}_a n_b + \tilde{g}_{ab}n^c\tilde{\nabla}_c\omega = \frac{1}{4}\tilde{g}_{ab}(\omega\tilde{g}^{cd}\tilde{\nabla}_c\tilde{\nabla}_d\Omega + 4n^c\tilde{\nabla}_c\omega) \tag{5.31}$$

where $n^c\tilde{\nabla}_c$ is the directional derivative along a null vector field on $\mathcal{J}^{\pm}$. We can choose an arbitrary ω on a sphere $u = \text{const}$ on $\mathcal{J}^+$ (or $v = \text{const}$ on $\mathcal{J}^-$) as an initial condition for the differential equation

$$n^c\tilde{\nabla}_c\omega = -\frac{1}{4}\omega\tilde{g}^{cd}\tilde{\nabla}_c\tilde{\nabla}_d\Omega$$

and then solve it along null generators of $\mathcal{J}^{\pm}$. The solution, extended to a neighborhood of $\mathcal{J}^{\pm}$, will provide a conformal gauge in which $\tilde{\nabla}'_a n'_b = 0$, at least as long as the solution for ω with the chosen initial values remains non-zero.

As the normal to the null surface $\mathcal{J}^{\pm}$ in $(\tilde{M}', \tilde{g}'_{ab})$ in the asymptotically flat case, n'^a can be used as the tangent vector field to the congruence of null curves defined by its trajectories along $\mathcal{J}^{\pm}$. In the conformal gauge with $B'_{ab} = \tilde{\nabla}'_a n'_b = 0$, this congruence is free of expansion, shear and rotation.

5.3.4 Asymptotic solution for asymptotically flat space-times

Finding a simple representation of the asymptotic metric is one application of the conformal gauge with a normal vector n^a on $\mathcal{J}^+$ making the congruence expansion, shear and rotation-free. Topologically, $\mathcal{J}^+$ is of the form $S^2 \times \mathbb{R}$. Any spherical cross-section S is called a *cut* and can be used to construct the general form of asymptotically flat space-times:

1. First, choose a second null hypersurface Σ_S through S not tangent to $\mathcal{J}^+$, such that its null generator ℓ^a is orthogonal to n^a as in Fig. 5.15, and such that both fields are cross-normalized: $\ell^a n_a = -1$ on S.
2. Extend ℓ^a to a vector field on all of $\mathcal{J}^+$ by Lie transport: $0 = n^a\nabla_a\ell^b = \mathcal{L}_n\ell^b$. (The second identity follows using $\nabla_a n_b = 0$, as per our conformal gauge on $\mathcal{J}^+$.) Since n^a is tangent to

affinely parameterized null geodesics (again using $\nabla_a n_b = 0$, in particular $n^a \nabla_a n_b = 0$), the cross-normalization condition $\ell^a n_a = -1$ chosen on S is preserved by this extension and holds on all of $\mathcal{J}^+$.

3. On every cut we write the induced metric in the form $\hat{h}_{ab} = g_{ab} + 2\ell_{(a}n_{b)}$ such that $\ell^a \hat{h}_{ab} = 0 = \hat{h}_{ab}n^a$. This induced metric is constant along $\mathcal{J}^+$: in addition to $\mathcal{L}_n n^a = 0$ (which is obvious) and $\mathcal{L}_n \ell^b = 0$ used above, we have $\mathcal{L}_n g_{ab} = \nabla_{(a}n_{b)} = 0$ and thus $\mathcal{L}_n \hat{h}_{ab} = 0$.
4. The condition $\nabla_a n_b = 0$ does not fix the conformal gauge completely. We can still change our conformal factor by $\Omega \mapsto \omega\Omega$ with $\omega \neq 0$ on $\mathcal{J}^+$ such that $n^c \nabla_c \omega = 0$. This freedom corresponds to the choice of initial values for the differential equation satisfied by ω. Thus, ω must be constant along the null direction of $\mathcal{J}^+$, but it can be freely specified, as long as it is non-zero, on an initial cut. In particular, on S the 2-dimensional Ricci scalar associated with $\hat{h}_{ab}$ will map to
$$^{(2)}\tilde{R} = \omega^2\, {}^{(2)}R + 2\omega \tilde{\nabla}^a \tilde{\nabla}_a \omega - 2(\tilde{\nabla}^a \omega)\tilde{\nabla}_a \omega \tag{5.32}$$
by the new conformal transformation. On a 2-dimensional manifold, there is always a conformal transformation bringing the space to constant-curvature form. In particular, we can choose ω so as to make the 2-metric equal $\hat{h}_{ab}\mathrm{d}x^a \mathrm{d}x^b = \mathrm{d}\vartheta^2 + \sin^2\vartheta \mathrm{d}\varphi^2$. This finally fixes the conformal gauge and makes every cut a round 2-sphere of unit radius.
5. The round 2-sphere contributions already give us the angular part of the metric; what remains to be done is to determine the form of the directions transversal to the cuts. For an explicit representation of the asymptotic space-time metric, we choose as the remaining coordinates the new conformal factor Ω in a neighborhood of $\mathcal{J}^+$, where the condition $\nabla_a \Omega \neq 0$ near $\Omega = 0$ indeed guarantees that Ω is a good coordinate, together with an affine parameter u along n^a, normalized such that $n^a \nabla_a u = 1$. Initially, ϑ and φ of the round 2-sphere cuts as well as u are only defined on $\mathcal{J}^+$, but we can transport them as functions into a neighborhood of the boundary by requiring that they have constant values along the null generators of Σ_S. The line element resulting in the neighborhood will depend on the choice made for Σ_S, but its asymptotic form will not. With all conditions imposed, the space-time metric near $\mathcal{J}^+$ in these coordinates is
$$\mathrm{d}\tilde{s}^2 \sim -2\mathrm{d}\Omega \mathrm{d}u + \mathrm{d}\vartheta^2 + \sin^2\vartheta \mathrm{d}\varphi^2 \,. \tag{5.33}$$
As one can directly verify, this indeed makes $\nabla_a \Omega$ and $\nabla_a u$ null, as well as $g^{ab}(\nabla_a \Omega)(\nabla_b u) = -1$ for the cross-normalization.

5.3.5 Asymptotic symmetries

The preceding construction has provided an asymptotic metric in a fixed conformal gauge, but not all the ingredients were free of choices. In particular, while there is no intrinsic restriction on the conformal factor other than that it be non-zero everywhere, the metric on an initial cut was conveniently but rather arbitrarily chosen to be the round 2-sphere metric of radius one. Moreover, the embedding of the round sphere as a cut of $\mathcal{J}^\pm$ depends on the coordinate u along $\mathcal{J}^\pm$ and on what one considers as its constant-level surfaces. This freedom gives rise to a specific form of coordinates combined with conformal transformations as asymptotic symmetries. These transformations are required only to leave the structure of $\mathcal{J}^{\pm}$ itself invariant, i.e. the 3-cylinder $S^2 \times \mathbb{R}$ endowed as a null surface with a "null metric" structure $(\hat{h}_{ab}\mathrm{d}x^a\mathrm{d}x^b, \mathrm{d}u^2)$.

On S, we had chosen a metric that can be expressed as

$$\mathrm{d}s^2 = \mathrm{d}\vartheta^2 + \sin^2\vartheta\,\mathrm{d}\varphi^2 = \frac{4}{(1+\zeta\bar{\zeta})^2}\mathrm{d}\zeta\,\mathrm{d}\bar{\zeta} \tag{5.34}$$

in complex coordinates $\zeta = \xi + i\eta = e^{i\varphi}\cot(\vartheta/2)$. (The bar denotes complex conjugation.) A coordinate transformation on S given by $\zeta \mapsto \zeta'(\zeta,\bar{\zeta})$ such that $\mathrm{d}\zeta' = A(\zeta,\bar{\zeta})\mathrm{d}\zeta + B(\zeta,\bar{\zeta})\mathrm{d}\bar{\zeta}$, $\mathrm{d}\bar{\zeta}' = \overline{B(\zeta,\bar{\zeta})}\mathrm{d}\zeta + \overline{A(\zeta,\bar{\zeta})}\mathrm{d}\bar{\zeta}$ is then conformal if and only if $A = 0$ or $B = 0$: in these two cases, and only in these, the new metric given by

$$\mathrm{d}\zeta'\mathrm{d}\bar{\zeta}' = A(\zeta,\bar{\zeta})\overline{B(\zeta,\bar{\zeta})}\mathrm{d}\zeta^2 + \left(|A(\zeta,\bar{\zeta})|^2 + |B(\zeta,\bar{\zeta})|^2\right)\mathrm{d}\zeta\,\mathrm{d}\bar{\zeta} + \overline{A(\zeta,\bar{\zeta})}B(\zeta,\bar{\zeta})\mathrm{d}\bar{\zeta}^2$$

is proportional to the original round 2-sphere metric in complex coordinates.

If $A = 0$, we must have

$$\begin{aligned}\mathrm{d}\zeta' = \mathrm{d}\xi' + i\,\mathrm{d}\eta' &= \left(\frac{\partial\xi'}{\partial\xi} + i\frac{\partial\eta'}{\partial\xi}\right)\mathrm{d}\xi + \left(\frac{\partial\xi'}{\partial\eta} + i\frac{\partial\eta'}{\partial\eta}\right)\mathrm{d}\eta \\ &= B(\zeta,\bar{\zeta})\mathrm{d}\bar{\zeta} = B(\zeta,\bar{\zeta})(\mathrm{d}\xi - i\,\mathrm{d}\eta)\end{aligned}$$

and thus

$$\frac{\partial\xi'}{\partial\xi} = \mathrm{Re}B = -\frac{\partial\eta'}{\partial\eta}\,, \qquad \frac{\partial\xi'}{\partial\eta} = \mathrm{Im}B = \frac{\partial\eta'}{\partial\xi}\,. \tag{5.35}$$

Such a transformation satisfying $\partial\zeta'/\partial\zeta = 0$ is called *antiholomorphic*.

Similarly, $B = 0$ implies $\mathrm{d}\zeta' = \mathrm{d}\xi' + i\,\mathrm{d}\eta' = A(\zeta,\bar{\zeta})(\mathrm{d}\xi - i\,\mathrm{d}\eta)$, such that

$$\frac{\partial\xi'}{\partial\xi} = \mathrm{Re}A = \frac{\partial\eta'}{\partial\eta}\,, \qquad \frac{\partial\xi'}{\partial\eta} = -\mathrm{Im}A = -\frac{\partial\eta'}{\partial\xi}\,. \tag{5.36}$$

Such a transformation satisfying $\partial\zeta'/\partial\bar{\zeta} = 0$ is called *holomorphic*.

If we restrict transformations to those that preserve orientation, we can only allow holomorphic ones of the form $\zeta \mapsto \zeta'(\zeta)$, i.e. $\mathrm{d}\zeta'/\mathrm{d}\bar{\zeta} = 0$, since complex conjugation $\xi + i\eta \mapsto \xi - i\eta$ changes orientation. This is not much of a restriction because every antiholomorphic transformation can be realized as a holomorphic one followed by complex conjugation. Then, any 1-to-1 holomorphic map on S^2 is given by a *fractional linear transformation*

$$\zeta'(\zeta) = \frac{a\zeta + b}{c\zeta + d} \quad \text{with complex numbers such that } ad - bc = 1\,. \tag{5.37}$$

This formula parameterizes all coordinate transformations for which we can compensate the change of metric on a cut of $\mathcal{J}^{\pm}$ by a conformal transformation $\Omega \mapsto \omega\Omega$.

A holomorphic transformation $\zeta'(\zeta)$ conformally transforms the 2-metric (5.34) on S to

$$\hat{h}'_{ab} = \left|\frac{\mathrm{d}\zeta'}{\mathrm{d}\zeta}\right|^2 \left(\frac{1+\zeta\bar{\zeta}}{1+\zeta'\bar{\zeta}'}\right)^2 \hat{h}_{ab} =: \omega(\zeta,\bar{\zeta})^2\hat{h}_{ab}\,. \tag{5.38}$$

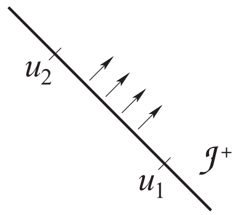

Fig. 5.16 Outgoing radiation as flow through $\mathcal{J}^+$, whose energy is given by the difference $M(u_2) - M(u_1)$.

We can thus compensate the metric change of the mapping by combining the holomorphic transformation of the coordinates with a conformal transformation of $\mathcal{J}^\pm$ given by ω. The whole structure $(\hat{h}_{ab}\mathrm{d}x^a\mathrm{d}x^b, \mathrm{d}u^2)$ is preserved if we also transform to $u'(u)$ such that $\mathrm{d}u' = \omega(\zeta, \bar{\zeta})\mathrm{d}u$. In particular, this will preserve the normalization of u by $n^a\nabla_a u = 1 = n'^a\nabla_a u'$. Along any null generator of $\mathcal{J}^\pm$, the most general coordinate transformation to perform is $u' = \omega(u + \alpha(\zeta, \bar{\zeta}))$. For non-constant α, constant-u cuts will be deformed; allowing for arbitrary smooth $\alpha(\zeta, \bar{\zeta})$ ensures independence of the specific cuts chosen.

Taken together, the transformations

$$(\zeta, u; \tilde{g}_{ab}) \mapsto (\zeta'(\zeta), u'(u, \zeta, \zeta'); \tilde{g}'_{ab}) \tag{5.39}$$

form a group representing asymptotic symmetries of $\mathcal{J}^\pm$: the *Bondi–Metzner–Sachs (BMS) group* introduced by Bondi et al. (1962) and Sachs (1962). Due to the freedom of a function $\alpha(\zeta, \zeta')$, it is infinite-dimensional and much bigger than the Poincaré symmetry group of Minkowski space that one might have expected in an asymptotically flat situation. A particular example of asymptotic symmetries is given by *super-translations* of the form $\zeta' = \zeta$, $u' = u + \alpha(\zeta, \zeta')$.

The size of the BMS group compared to the Poincaré group is clearly a consequence of the conformal transformations allowed to compensate changes of the metric under a diffeomorphism. Nevertheless, the increase in size is quite surprising, and it happens only for asymptotically flat space-times. In asymptotic de Sitter or anti de Sitter space-times, asymptotic symmetries do not differ from the de Sitter group O(4,1) and the anti de Sitter group O(3,2), respectively.

Asymptotic symmetries allow the introduction of conserved quantities such as asymptotic expressions for energy, momentum or angular momentum and their fluxes which would be associated with asymptotic translations and rotations, respectively. An example is the so-called Bondi mass $M(u)$, see Fig. 5.16, which is a function on $\mathcal{J}^+$ such that $M(u_2) - M(u_1)$ is the outgoing energy flux between u_1 and u_2. However, the BMS group does not have a distinguished subgroup isomorphic to the Poincaré group, and so there is no straightforward implementation of conserved quantities. A long series of developments to find asymptotic expressions for energy and other conserved quantities in the radiative regime has been made, e.g. an incomplete list given by Komar (1959), Geroch and Winicour (1981), Ashtekar

(1981), and based on canonical methods especially by Ashtekar and Streubel (1981), Iyer and Wald (1994) and Wald and Zoupas (2000).

5.3.6 Spatial infinity

Before radiation can carry away energy through $\mathcal{J}^+$, the original mass contained in space-time is computationally accessible at spatial infinity i^0. In stationary situations, this describes the mass of compact objects. The space-time metric near i^0 in an asymptotically flat space-time cannot be arbitrary but must allow asymptotically Cartesian coordinates x^μ such that it takes the form

$$g_{\mu\nu} \sim \delta_{\mu\nu} + O(r^{-1}), \qquad \partial_\rho g_{\mu\nu} \sim O(r^{-2}) \tag{5.40}$$

where r is the radius as a function of the spatial Cartesian coordinates. Energy should then equal the Hamiltonian of the canonical formulation, but as we have seen, its bulk part is the Hamiltonian constraint and vanishes for every solution. Non-zero energy can result only if there is a remaining boundary term that does not vanish for the given fall-off conditions, as already seen in Chapter 3.3.2 for quasilocal quantities associated with finite boundaries. In the asymptotic context, energy and other conserved quantities based on boundary terms have been derived by Regge and Teitelboim (1974).

In terms of the canonical variables, the fall-off conditions follow from the relation with the space-time metric:

$$h_{ab} \sim \delta_{ab} + O(r^{-1}), \qquad p^{ab} \sim O(r^{-2}) \tag{5.41}$$

$$N \sim 1 + O(r^{-1}), \qquad \partial_a N \sim O(r^{-2}) \tag{5.42}$$

$$N^a \sim O(r^{-1}), \qquad \partial_b N^a \sim O(r^{-2}). \tag{5.43}$$

From the variations of the action as in Chapter 4.2.2, we obtain several boundary terms at a fixed 2-sphere $S_r : r = \text{const}$ with spatial normal $r_a = (\mathrm{d}r)_a$:

$$2\int_{S_r} \mathrm{d}^2 y r_a p^{ab} \delta N_b \sim r^2 O(r^{-3}) = 0$$

$$2\int_{S_r} \mathrm{d}^2 y r_a p^{cg} N^a \delta h_{cg} \sim r^2 O(r^{-4}) = 0$$

$$\frac{1}{16\pi G}\int_{S_r} \mathrm{d}^2 y r_a N (h^{ac}h^{de} - h^{ae}h^{cd}) D_c \delta h_{de} \sim r^2 O(r^{-2}) \neq 0$$

$$\frac{1}{16\pi G}\int_{S_r} \mathrm{d}^2 y r_c (D_a N)(h^{ac}h^{de} - h^{ae}h^{cd}) \delta h_{de} \sim r^2 O(r^{-3}) = 0\,.$$

This gives rise to only one non-vanishing asymptotic boundary term in the variation

$$\begin{aligned}
\delta\int_{\Sigma} \mathrm{d}^3x N C_{\mathrm{grav}} &= \lim_{r\to\infty}\int_{S_r} \mathrm{d}^2y r_a N(h^{ac}h^{de} - h^{ae}h^{cd}) D_c\delta h_{de} \\
&\sim \lim_{r\to\infty}\int_{S_r} \mathrm{d}^2y r_a \delta^{ab}\delta^{cd}(\bar{D}_b\delta h_{cd} - \bar{D}_c\delta h_{bd}) \\
&\sim \lim_{r\to\infty}\delta\left(\int_{S_r} \mathrm{d}^2y r_a \delta^{ab}\delta^{cd}(\bar{D}_b h_{cd} - \bar{D}_c h_{bd})\right) \\
&=: -16\pi G\delta E_{\mathrm{ADM}}[h_{ab}]
\end{aligned} \tag{5.44}$$

where $\bar{D}_a$ denotes the asymptotically Cartesian spatial derivative operator which is independent of h_{ab}. Indices are explicitly raised or lowered with the asymptotically Cartesian metric δ_{ab}.

The total Hamiltonian must have vanishing variation once field equations are satisfied, which can be the case only if $E_{\mathrm{ADM}}[h_{ab}]$ is added as a boundary term to cancel the variation (5.44) of the bulk term. Thus,

$$H = \int_{\Sigma} \mathrm{d}^3x(N C_{\mathrm{grav}} + N^a C_a^{\mathrm{grav}}) + E_{\mathrm{ADM}}[h_{ab}] \tag{5.45}$$

is the total Hamiltonian. When field equations are satisfied, the constraints $C_{\mathrm{grav}} = 0 = C_a^{\mathrm{grav}}$ vanish such that the Hamiltonian on the space of solutions takes the value of the *ADM energy*

$$E_{\mathrm{ADM}} = -\frac{1}{16\pi G}\lim_{r\to\infty}\int_{S_r} \mathrm{d}^2y r_a \delta^{ab}\delta^{cd}(\bar{D}_b h_{cd} - \bar{D}_c h_{bd})\,. \tag{5.46}$$

Hawking and Horowitz (1996) have shown that E_{ADM} agrees with the asymptotic limit of the Brown–York quasilocal energy on spheres.

Momentum and angular momentum can be derived from the Nöther theorem, according to which the change of the action under a symmetry provides a conserved charge. Here, we consider coordinate changes $x^\mu \mapsto x^\mu + \xi^\mu(x)$ under which fields change by their Lie derivative along ξ^a. The ADM energy then remains invariant because we have $\mathcal{L}_\xi h_{ab} = 2\bar{D}_{(a}\xi_{b)}$ asymptotically, and $\bar{D}_b\bar{D}_a\xi^b - \bar{D}_a\bar{D}_b\xi^b = 0$ for the asymptotically flat metric.

The remaining terms in the Lagrangian required to be invariant are those in $\int \mathrm{d}^3x(p^{ab}\dot{h}_{ab} - N C_{\mathrm{grav}} - N^a C_a^{\mathrm{grav}})$, which change by an integrated divergence $\int \mathrm{d}^3x D_c\left((p^{ab}\dot{h}_{ab} - N C_{\mathrm{grav}} - N^a C_a^{\mathrm{grav}})\xi^c\right)$ (duly taking into account the Lie derivative of a scalar density). If the constraints are satisfied, only the boundary term $\int \mathrm{d}^3x D_c(p^{ab}\dot{h}_{ab}\xi^c) = \lim_{r\to\infty}\int_{S_r} \mathrm{d}^2y r_c\xi^c p^{ab}\dot{h}_{ab}$ must be required to vanish for an asymptotic symmetry. This provides some restrictions on vector fields ξ^a that can be asymptotic symmetries of the action. For any vector field for which the boundary term vanishes, Nöther's theorem provides us with a conserved quantity because we have

$$0 = \delta S = \int \mathrm{d}^3x p^{ab}\delta h_{ab}|_{t_1}^{t_2} \tag{5.47}$$

for the boundary term resulting from integrating by parts in time. The conserved charge is thus

$$\int \mathrm{d}^3 x p^{ab} \delta_\xi h_{ab} = 2 \int \mathrm{d}^3 x p^{ab} D_{(a}\xi_{b)} = 2 \int \mathrm{d}^3 x \left(D_a(p^{ab}\xi_b) - \xi_b D_a p^{ab} \right)$$
$$= 2 \lim_{r\to\infty} \int_{S_r} \mathrm{d}^2 y r_a p^{ab} \xi_b \tag{5.48}$$

for any ξ^a such that $r_a\xi^a = O(1)$ on S_r with its co-normal $r_a = (\mathrm{d}r)_a$. This formula agrees with the asymptotic limit of the Brown–York momentum in (3.71) for translations along the shift $\xi^b = N^b/N$, using the fact that the reference momentum $\bar{p}^{ab}$ drops off to zero fast enough.

Examples for vector fields satisfying this condition are asymptotic translations and asymptotic rotations, providing momentum and angular momentum as conserved quantities. For an asymptotic translation, $\xi^a = -\epsilon^a$ is constant and the *ADM momentum*

$$P^a = -2 \lim_{r\to\infty} \int_{S_r} \mathrm{d}^2 y r_b p^{ab} \tag{5.49}$$

results. For an asymptotic rotation $\xi_a = -\epsilon_{abc}\phi^b x^c$ with a constant rotation vector ϕ^a (whose length gives the rotation angle) we have the *angular momentum*

$$L_b = -2 \lim_{r\to\infty} \int_{S_r} \mathrm{d}^2 y r_a \epsilon_{bcd} p^{ac} x^d \,. \tag{5.50}$$

Note, however, that there is arbitrariness in this definition, since angular momentum refers to an origin of rotations in Cartesian coordinates which may lie outside the asymptotic Cartesian region. Different choices of coordinates (different asymptotic observers and their synchronizations) can give rise to different values of angular momentum.

For other notions of energy and angular momentum in general relativity, see the review by Jaramillo and Gourgoulhon (2010).

5.4 Matching of solutions

Matching techniques of different space-time regions to result in one global solution provide the means to construct conformal diagrams of complicated space-times from those of simpler ones.

The Schwarzschild space-time, for instance, is an explicit solution illustrating several general properties of black holes such as trapped regions, horizons and singularities. But it describes neither the formation of the black hole in a collapse process nor dynamical properties after the black hole has formed. It is thus of interest to construct models of collapsing matter that display the general behavior of trapped surfaces and the consequences of horizons and singularities.

To model a collapsing star, we require a distribution of matter in its interior with vacuum space-time outside the star's surface. The whole space-time has to obey Einstein's equation,

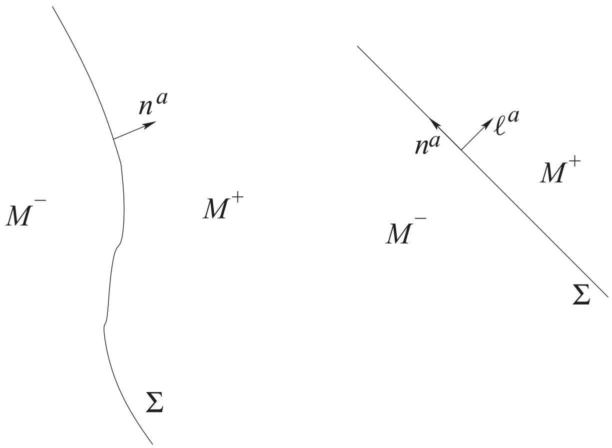

Fig. 5.17 The matching surface Σ splits space-time into components $M^{\pm}$, with the normal vector n^a of Σ (or the cross-normalized vector ℓ^a in the null case) pointing toward M^+.

for whose solutions several examples are easily available separately in the presence of matter as well as in vacuum. However, we must also ensure that the equation is satisfied for fields residing on the star surface, separating the well-known regions. This can be done by matching interior and vacuum solutions at this boundary, imposing conditions for limiting values of the fields when the surface is approached from both sides as part of a complete solution.

Idealizing the surface as a sharp transition from matter to vacuum requires some care in discussing the fields. At the surface, the energy-momentum tensor T_{ab} is discontinuous. On the other hand, not all geometric quantities can be discontinuous if the Einstein tensor G_{ab} and the derivatives it contains are to be regular. This observation provides matching conditions for suitable space-time metrics in the two bordering regions. If these conditions can be satisfied for given choices of inside and outside metrics, a valid model of a star is obtained. General discussions of matching in different situations have been provided by Clarke and Dray (1987) and Mars and Senovilla (1993), based on seminal work by Israel (1966).

In this setting, as sketched in Fig. 5.17, we have the matching surface Σ bounding two different space-time regions $(M^{\pm}, g^{\pm}_{ab})$. Since both regions overlap at Σ, there must be a map $\partial M^- \to \partial M^+$ linking the boundaries of M^+ and M^-, which then provides an identification of the two tangent spaces induced on Σ by viewing the surface as the boundary of M^+ and M^-, respectively, identified as one surface by the mapping. In order to complete the identification of the full tangent spaces of $M^{\pm}$ at Σ, we also identify the normals $n^a_{\pm} =: n^a$ of Σ. (We choose n^a_+ to be inward-pointing and n^a_- outward-pointing to be specific.) If the matching surface is null, a case we will discuss separately, n^a is tangent to the surface, and we choose a normal ℓ^a cross-normalized with n^a, $\ell^a n_a = -1$, pointing into M^+.

Example 5.4 (Null surfaces)
Null surfaces and their normals have already been used several times. At this stage, some of the differences to non-null surfaces will become important. Consider the example of the null surface $\Sigma \colon x - t = \mathrm{const}$ *in Minkowski space. We directly obtain the co-normal* $n_a = (\mathrm{d}(x-t))_a = (\mathrm{d}x)_a - (\mathrm{d}t)_a$*, and one can easily verify that it is co-normal to the surface:* $n_a v^a = 0$ *for all* v^a *tangent to the surface. Raising the index, we have* $n^a = (\partial/\partial x)^a + (\partial/\partial t)^a$*, a tangent vector to the surface, since* $n_a n^a = 0$ *with* n^a *null. Indeed,* $n^a \nabla_a (x-t) = 0$ *such that displacements along* n^a *stay on the surface* Σ*. A transversal vector to the null surface is obtained not by raising the index of the co-normal, but by determining the null vector cross-normalized with* n_a*. Here,* $\ell^a = \frac{1}{2}(-(\partial/\partial x)^a + (\partial/\partial t)^a)$ *is null, transversal to* Σ *(*$\ell^a \nabla_a (x-t) \neq 0$*), and satisfies* $\ell^a n_a = -1$*. Both* n^a *and* ℓ^a *are future-pointing, which is ensured by the cross-normalization to a negative number.*

Both regions, (M^+, g^+_{ab}) and (M^-, g^-_{ab}), induce metrics $h^{\pm}_{ab}$ on Σ, which must be identical for a valid matching; otherwise, the extrinsic curvature $K_{ab} = \frac{1}{2}\mathcal{L}_n h_{ab}$ of Σ in the resulting space-time would be singular on Σ, and, with the curvature relations from Chapter 3.2.3, the Ricci tensor R_{ab} would have a singularity. Since the metric is regular, even though discontinuous, the Ricci scalar $g^{ab} R_{ab}$ would diverge, implying a curvature singularity.

If the induced metrics agree, on the other hand, all tangential derivatives of the space-time metric are continuous: for any v^c with $v^c n_c = 0$ we have $v^c \Delta(\partial_c g_{ab}) = 0$ where we use the notation

$$\Delta f(x) = \lim_{M^+ \ni y \to x} f(y) - \lim_{M^- \ni y \to x} f(y) \tag{5.51}$$

for the discontinuity of a function f. The space co-normal to Σ is spanned by n^+_a, and so derivatives of the metric must have a discontinuity of the general form $\Delta(\partial_c g_{ab}) = n_c \gamma_{ab}$ with a symmetric (non-tensorial) object γ_{ab} that measures the discontinuity of $\partial_c g_{ab}$. From this, a discontinuity

$$\Delta\Gamma^a_{bc} = \frac{1}{2}\Delta\left(g^{ad}(\partial_b g_{dc} + \partial_c g_{bd} - \partial_d g_{bc})\right) = \frac{1}{2}(n_b \gamma^a_c + n_c \gamma^a_b - n^a \gamma_{bc}) \tag{5.52}$$

of the connection coefficients follows: Γ^a_{bc} is the sum of a continuous distribution and $\theta_\Sigma \Delta\Gamma^a_{bc}$ with the surface step function θ_Σ taking the values $\pm\frac{1}{2}$ in $M^\pm$.

Example 5.5 (Surface step and delta functions)
We use the step function θ_Σ *with values* $\pm\frac{1}{2}$ *in the manifolds* $M^\pm$ *with common boundary* Σ *in order to describe discontinuities across* Σ*. Its derivatives along directions not tangent to* Σ *result in Dirac-type delta distributions* δ_Σ*, vanishing in* $M^\pm \backslash \Sigma$ *and satisfying* $\int \delta_\Sigma \mathrm{d}n = 1$ *when integrated along a curve* $n(t)$ *from* M^- *to* M^+*, normal to* Σ *at the intersection point. The tangent vector of the curve where it intersects* Σ *is thus* n^a *if* Σ *is non-null, and* ℓ^a *if it is null.*

While tangential derivatives of θ_Σ *vanish,* $v^c \nabla_c \theta_\Sigma = 0$ *for all* v^a *tangent to* Σ*, normal derivatives result in delta distributions. Again using the fact that the co-normal space to* Σ *is spanned by* n_a*, we can write* $\nabla_c \theta_\Sigma \propto n_c \delta_\Sigma$*. To determine the coefficient, we integrate*

$\int n^a \nabla_a \theta_\Sigma \mathrm{d}n = \theta_\Sigma^+ - \theta_\Sigma^- = 1$ and compare with $\int n^a n_a \delta_\Sigma \mathrm{d}n = -1$ for a timelike surface and $\int n^a n_a \delta_\Sigma \mathrm{d}n = 1$ for a spacelike one. For a null surface, ℓ^a is used instead of n^a as the tangent to the normal curve (while the co-normal n_a remains in $\nabla_c \theta_\Sigma \propto n_c \delta_\Sigma$), also resulting in $\int \ell^a n_a \delta_\Sigma \mathrm{d}n = -1$ with cross-normalization. Comparison then fixes $\nabla_c \theta_\Sigma = -n_c \delta_\Sigma$ for timelike and null surfaces, with a plus sign instead of the minus for a spacelike surface.

Then, also R_{ab} will be discontinuous or could even be singular of δ-function form: $R_{ab} = R_{ab}^{\text{non-sing}} + R_{ab}^{\text{sing}} \delta_\Sigma$. The coefficient of the δ-function arises from normal derivatives in the general expression

$$R_{ab} = \nabla_c \Gamma^c_{ab} - \nabla_b \Gamma^c_{ca} + \Gamma^d_{ab} \Gamma^c_{dc} - \Gamma^d_{cb} \Gamma^c_{da}$$

of the Christoffel coefficients and takes the form

$$\begin{aligned} R_{ab}^{\text{sing}} &= n_c \Delta \Gamma^c_{ab} - n_b \Delta \Gamma^c_{ac} \\ &= \frac{1}{2}(n_a n_c \gamma_b^c + n_c n_b \gamma_a^c - n_c n^c \gamma_{ab} - n_b n_a \gamma_c^c - n_b n_c \gamma_a^c + n_b n^c \gamma_{ac}) \\ &= n^c n_{(b} \gamma_{a)c} - \frac{1}{2} n^c n_c \gamma_{ab} - \frac{1}{2} n_a n_b \gamma_c^c \,. \end{aligned}$$

The Ricci scalar thus has a δ-function coefficient

$$R^{\text{sing}} = n^a n^b \gamma_{ab} - n^a n_a \gamma_b^b \tag{5.53}$$

and the Einstein tensor one of the form

$$\begin{aligned} G_{ab}^{\text{sing}} &= R_{ab}^{\text{sing}} - \frac{1}{2} R^{\text{sing}} g_{ab} \\ &= n^c n_{(b} \gamma_{a)c} - \frac{1}{2} n^c n_c \gamma_{ab} - \frac{1}{2} n_a n_b \gamma_c^c - \frac{1}{2} g_{ab} (n^c n^d \gamma_{cd} - n^c n_c \gamma_d^d) \,. \end{aligned} \tag{5.54}$$

These equations are valid for any value of $n^c n_c$, including the null case in which some terms drop out. Further considerations and the precise matching conditions, however, depend on whether the normal is null.

5.4.1 Non-null matching surface

If the surface is not null, we can normalize the normal to $n^a n_a = \pm 1$ and explicitly write the induced metric on the matching surface Σ as $h_{ab} = g_{ab} \pm n_a n_b$. The induced metric is continuous, producing the same value, irrespective of whether we induce it from M^+ or

M^-. Extrinsic curvature is regular but discontinuous:

$$\begin{aligned}\Delta K_{ab} &= \Delta(h_a^c \nabla_c n_b) = -\Delta(\Gamma^d_{cb} n_d h_a^c) \\ &= -\frac{1}{2}(n_c n_d \gamma_b^d + n_b n_d \gamma_c^d \pm \gamma_{cb})(g_a^c \pm n^c n_a) \\ &= -\frac{1}{2}(2 n_d n_{(a} \gamma_{b)}^d \pm \gamma_{ab} - n_d n_a \gamma_b^d \pm n_a n_b n^c n^d \gamma_{cd} + n^c n_a \gamma_{cb}) \\ &= -n_d n_{(a} \gamma_{b)}^d \mp \frac{1}{2}\gamma_{ab} \mp \frac{1}{2} n_a n_b n^c n^d \gamma_{cd} \end{aligned} \tag{5.55}$$

and

$$\Delta K_a^a = -\frac{1}{2}(n^a n^b \gamma_{ab} \pm \gamma_a^a)\,. \tag{5.56}$$

If we compare this with (5.54), we see that

$$G_{ab}^{\mathrm{sing}} = -\Delta K_{ab} + h_{ab} \Delta K_c^c \tag{5.57}$$

which shows that the Einstein tensor is non-singular if and only if K_{ab} is continuous.

For a regular stress-energy tensor, the Einstein tensor must be regular, which provides the *junction conditions* $K_{ab}^+ = K_{ab}^-$ in addition to $h_{ab}^+ = h_{ab}^-$. If the energy-momentum tensor has a δ-function singularity, which would mean that there is a surface layer of energy-momentum on Σ, the Einstein tensor is allowed to have a singular part, required to agree with that of the energy-momentum tensor: $G_{ab}^{\mathrm{sing}} = 8\pi G T_{ab}^{\mathrm{sing}}$. Via the resulting discontinuity in K_{ab}, fixed in terms of T_{ab}^{sing}, the matching is again determined.

Practically, to see whether one can consistently match two space-times along a surface, one makes a choice of how to embed the matching surface in each of them, and then computes the induced metrics and extrinsic curvatures as seen from both space-times. Finally, one checks whether they can be made to agree. This is most interesting in the case of a timelike rather than spacelike matching surface, for the spacelike matching is nothing but a rederivation of the initial value problem of Chapter 3.4: junction conditions for spacelike Σ can be interpreted as saying that on Σ we must have initial conditions for M^+ in agreement with the final geometry as seen from M^- (if n^a is future pointing to M^+). The timelike case is, however, new and provides interesting possibilities for constructing new solutions. A detailed example is provided after the discussion of null matching surfaces.

5.4.2 Null matching surface

If $n^a n_a = 0$, we have a second null vector ℓ^a and cross-normalized with n_a: $\ell^a n_a = -1$. Instead of a spatial metric on all of Σ, we induce a spatial metric q_{ab} on cross-sections of Σ normal to both n^a and ℓ^a by $g_{ab} = q_{ab} - 2\ell_{(a} n_{b)}$. In this case, instead of extrinsic curvature, various derivatives of both null normals appear. The singular part of the Einstein tensor becomes

$$
\begin{aligned}
G_{ab}^{\text{sing}} &= n^c n_{(b}\gamma_{a)c} - \frac{1}{2}n_a n_b \gamma_c^c - \frac{1}{2}g_{ab}n^c n^d q_{cd} \\
&= n^c n_{(b}\gamma_{a)c} - \frac{1}{2}n_a n_b \gamma_{cd}(q^{cd} - 2\ell^{(c}n^{d)}) - \frac{1}{2}q_{ab}n^c n^d \gamma_{cd} + \ell_{(a}n_{b)}n^c n^d \gamma_{cd} \\
&= -\frac{1}{2}n_a n_b \gamma_{cd} q^{cd} + n_a n_b \ell^c n^d \gamma_{cd} + n^c n_{(b}\gamma_{a)c} + \ell_{(a}n_{b)}n^c n^d \gamma_{cd} - \frac{1}{2}q_{ab}n^c n^d \gamma_{cd} \\
&= n_a n_b \Delta\theta_\ell - 2\Delta\eta_{(a}n_{b)} - \Delta\omega q_{ab}
\end{aligned}
\tag{5.58}
$$

where we used $n_a n_b \ell^c n^d \gamma_{cd} + n^c n_{(b}\gamma_{a)c} + \ell_{(a}n_{b)}n^c n^d \gamma_{cd} = n^d \gamma_{cd} q_{(a}^c n_{b)}$ and introduced $\theta_\ell := q^{cd}\nabla_c \ell_d$ (the expansion of the null congruence defined by ℓ^a), $\eta_a := q_a^c n^d \nabla_c \ell_d$ and $\omega := \ell_a n^c \nabla_c n^a$. They have discontinuities

$$
\begin{aligned}
\Delta\theta_\ell &= -q^{cd}\Delta\Gamma_{cd}^e \ell_e = -\frac{1}{2}q^{cd}\gamma_{cd} \\
\Delta\eta_a &= -q_a^c n^d \Delta\Gamma_{cd}^e \ell_e = -\frac{1}{2}q_a^c n^d \gamma_{cd} \\
\Delta\omega &= \ell_a n^c \Delta\Gamma_{cb}^a n^b = \frac{1}{2}n^b n^c \gamma_{bc}
\end{aligned}
$$

which justifies our use of these quantities in G_{ab}^{sing}. Thus, for continuous θ_ℓ, η_a and ω the Einstein tensor is non-singular on the matching surface. There are only four conditions rather than six for the symmetric K_{ab} in the non-null case; null-junction conditions are thus weaker. In particular, the full Riemann tensor $R_{abc}{}^d$ can be singular on a null matching surface even if the Einstein tensor is non-singular, i.e. without a matter source of a surface layer. Physically, this singular situation corresponds to a gravitational shock wave.

5.4.3 Oppenheimer–Snyder model

Oppenheimer and Snyder (1939) model a collapsing homogeneous star surrounded by vacuum by matching an FLRW solution as the interior to a Schwarzschild exterior. As the collection of world-lines for all particles on the star's surface, the matching surface Σ should be timelike.

Our interior metric is now given by

$$
\mathrm{d}s_-^2 = -\mathrm{d}\tau^2 + X(\tau,r)^2\mathrm{d}r^2 + Y(\tau,r)^2\mathrm{d}\Omega^2 \tag{5.59}
$$

where

$$
X(\tau,r) = \frac{a(\tau)}{\sqrt{1-kr^2}}\,, \qquad Y(\tau,r) = a(\tau)r\,. \tag{5.60}
$$

In contrast to a cosmological model, we do not use this line element for all of space-time, but, as in Fig. 5.18, cut off the manifold at a value $r = R$ to define the matching surface Σ as a boundary of M^-. Its unit normal vector is $n_-^a = X^{-1}(\partial/\partial r)^a$ and it inherits an induced

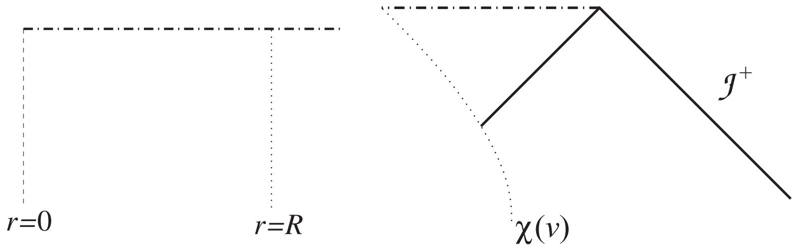

Fig. 5.18 Oppenheimer–Snyder model: a portion of an FLRW space-time between $r = 0$ and $r = R$ is matched to a portion of the Schwarzschild space-time along $\chi(v)$.

metric ($r = \text{const}$) $h^-_{ab}\mathrm{d}x^a\mathrm{d}x^b = -\mathrm{d}\tau^2 + Y^2\mathrm{d}\Omega^2$. This gives us the extrinsic curvature

$$K^-_{ab} = \frac{1}{2}\mathcal{L}_n h^-_{ab} = \frac{1}{2}(n^c_-\partial_c h^-_{ab} + h^-_{cb}\partial_a n^c_- + h^-_{ac}\partial_b n^c_-) \tag{5.61}$$

with components

$$K^-_{\tau\tau} = 0 = K^-_{\vartheta\tau}\,, \qquad K^-_{\vartheta\vartheta} = \frac{1}{2}n^r_-\partial_r h^-_{\vartheta\vartheta} = \frac{Y\partial_r Y}{X}$$

and similar components with φ instead of ϑ.

For the exterior, we use the Schwarzschild space-time with a metric in Eddington–Finkelstein form:

$$\mathrm{d}s^2_+ = -\left(1 - \frac{2GM}{\chi}\right)\mathrm{d}v^2 + 2\mathrm{d}v\mathrm{d}\chi + \chi^2\mathrm{d}\Omega^2\,. \tag{5.62}$$

This coordinate system allows us to cross the horizon, if necessary, without changing the chart. We only take a portion for $\chi > \chi(v)$ as an exterior region sketched in Fig. 5.18, to be matched to the FLRW interior. Since the precise relationship between the regions is unknown so far, we distinguish the coordinates (v, χ) from (τ, r) used for the interior. The relation $v(\tau)$, $\chi(\tau)$ at $r = R$, once derived, will determine the form of the matching surface in M^+. We have, however, already identified the angles in $\mathrm{d}\Omega$, as we are free to do, since both metrics are spherically symmetric.

On a general matching surface in M^+, parameterized by $(v(\tau), \chi(\tau))$ in terms of a parameter τ, later to be identified with the FLRW proper time, we have the induced metric

$$h^+_{ab}\mathrm{d}x^a\mathrm{d}x^b = -\left(\left(1 - \frac{2GM}{\chi}\right)\dot{v}^2 - 2\dot{v}\dot{\chi}\right)\mathrm{d}\tau^2 + \chi^2\mathrm{d}\Omega^2$$

where the dot denotes differentiation by τ. (Unnormalized) normal vector components obey

$$m^+_a = \nabla_a v - \frac{\dot{v}}{\dot{\chi}}\nabla_a\chi$$

obtained from $m^+_a\mathrm{d}x^a = \mathrm{d}v - \frac{\partial v}{\partial\chi}\mathrm{d}\chi = \mathrm{d}v - \frac{\dot{v}}{\dot{\chi}}\mathrm{d}\chi$ as the co-normal vector to the surface $v - v(\chi) = \text{const}$. (This holds locally where $\chi(\tau)$ is invertible such that $t(\chi)$ can be inserted in $v(\tau)$ to obtain $v(\chi)$.) With the norm

$$g^+_{ab}m^a_+m^b_+ = 1 - \frac{2GM}{\chi} - 2\frac{\dot\chi}{\dot v}$$

the unit normal vector components are

$$n^v_+ = \frac{1}{\sqrt{1-2GM/\chi - 2\dot\chi/\dot v}}\,, \qquad n^\chi_+ = \frac{1-2GM/\chi - \dot\chi/\dot v}{\sqrt{1-2GM/\chi-2\dot\chi/\dot v}}\,.$$

This gives us the relevant components of extrinsic curvature in

$$\left(K^+_{vv} + 2K^+_{v\chi}\frac{\mathrm{d}\chi}{\mathrm{d}v} + K^+_{\chi\chi}\left(\frac{\mathrm{d}\chi}{\mathrm{d}v}\right)^2\right)\mathrm{d}v^2$$

which we are not going to need explicitly, and

$$K^+_{\vartheta\vartheta} = \frac{1}{2}n^\chi_+\partial_\chi h^+_{\vartheta\vartheta} = \chi\,\frac{1-2GM/\chi-\dot\chi/\dot v}{\sqrt{1-2GM/\chi-2\dot\chi/\dot v}}\,.$$

With these preparations, we formulate the junction conditions:

$h^+_{\vartheta\vartheta} = h^-_{\vartheta\vartheta}$: we obtain $\chi(\tau) = Y(\tau, r) = a(\tau)R$ which for a given interior $a(\tau)$ determines χ on the exterior matching surface up to the free parameter R.

$h^+_{\tau\tau} = h^-_{\tau\tau}$: the equation $(1-2GM/\chi)\,\dot v^2 - 2\dot v\dot\chi = 1$ provides a differential equation

$$\left(\frac{\mathrm{d}\chi}{\mathrm{d}v}\right)^2 = \frac{\dot\chi^2}{\dot v^2} = \dot a^2R^2\left(1-\frac{2GM}{\chi} - 2\frac{\mathrm{d}\chi}{\mathrm{d}v}\right) \tag{5.63}$$

for $\chi(v)$. The relation between the exterior coordinates is now determined (up to the value of R), fixing the matching surface in M^+.

$K^+_{\tau\tau} = K^-_{\tau\tau}$: another differential equation

$$K^+_{vv} + 2K^+_{v\chi}\frac{\mathrm{d}\chi}{\mathrm{d}v} + K^+_{\chi\chi}\left(\frac{\mathrm{d}\chi}{\mathrm{d}v}\right)^2 = 0$$

results for $\chi(v)$, which turns out to be automatically satisfied given the other junction conditions. Together with the $h_{\tau\tau}$-matching, we have fixed $\chi(\tau)$ and $v(\tau)$ consistently. After this matching of the boundaries of M^- and M^+ as a single matching surface Σ, no free function is left and the only remaining parameter is R.

$K^+_{\vartheta\vartheta} = K^-_{\vartheta\vartheta}$: a final equation

$$\chi\,\frac{1-2GM/\chi-\mathrm{d}\chi/\mathrm{d}v}{\sqrt{1-2GM/\chi-2\mathrm{d}\chi/\mathrm{d}v}} = \frac{YY'}{X} \tag{5.64}$$

remains to be satisfied on the whole matching surface. Using $\chi = Y$ from the $h_{\vartheta\vartheta}$-matching, this can be written as

$$
\begin{aligned}
&\left(\frac{Y'}{X}\right)^2 \left(1 - \frac{2GM}{\chi} - 2\frac{\mathrm{d}\chi}{\mathrm{d}v}\right) \\
&= \left(1 - \frac{2GM}{\chi} - \frac{\mathrm{d}\chi}{\mathrm{d}v}\right)^2 \\
&= \left(1 - \frac{2GM}{Y}\right)^2 - 2\frac{\mathrm{d}\chi}{\mathrm{d}v}\left(1 - \frac{2GM}{Y}\right) + \left(\frac{\mathrm{d}\chi}{\mathrm{d}v}\right)^2 \\
&= \left(1 - \frac{2GM}{Y}\right)^2 - 2\frac{\mathrm{d}\chi}{\mathrm{d}v}\left(1 - \frac{2GM}{Y}\right) + \dot{Y}^2\left(1 - \frac{2GM}{\chi} - 2\frac{\mathrm{d}\chi}{\mathrm{d}v}\right) \\
&= \left(1 - \frac{2GM}{\chi} - 2\frac{\mathrm{d}\chi}{\mathrm{d}v}\right)\left(1 - \frac{2GM}{Y} + \dot{Y}^2\right)
\end{aligned}
$$

where we used the $h_{\tau\tau}$-matching in the next-to-last step. This equation can then be solved for the mass

$$
\begin{aligned}
GM &= \frac{1}{2}Y\left(1 + \dot{Y}^2 - \left(\frac{Y'}{X}\right)^2\right) = \frac{1}{2}aR\left(1 + r^2\dot{a}^2 - (1 - kR^2)\right) \\
&= \frac{1}{2}R^3 a(\dot{a}^2 + k)
\end{aligned}
$$

using that $X = a/\sqrt{1 - kR^2}$ and $Y = aR$ at the matching surface. This expression, as the mass parameter of the Schwarzschild solution, must be time independent, a condition which restricts the allowed matter content of the FLRW interior. The equation can be reformulated as

$$
\left(\frac{\dot{a}}{a}\right)^2 + \frac{k}{a^2} = \frac{2GM}{a^3R^3} = \frac{8\pi G}{3}\rho\,. \tag{5.65}
$$

By comparison with the Friedmann equation, pressure is required to vanish in the star in order for the energy density to behave like a^{-3}. For zero pressure, we then have

$$
M = \frac{4\pi}{3}R^3 a^3 \rho = V\rho \tag{5.66}
$$

which relates the star's mass to its volume and energy density ρ.

Now, all junction conditions are satisfied and we have a combined space-time solving Einstein's equation everywhere. The derivation shows that the existence of a consistent matching is non-trivial: there were several junction equations which made use of detailed properties of the interior and exterior metrics. In particular, an FLRW model with non-zero pressure cannot be matched to a Schwarzschild exterior. Physically, the non-zero pressure at the surface would cause matter to be ejected, in conflict with a vacuum exterior. (To model this situation, one could, for instance, use a Vaidya space-time or a generalized version of it, where the coordinate dependence $M(v)$ or $M(v, \chi)$ allows more freedom to satisfy the matching equations; see Exercise 5.12. The Vaidya space-time itself will be discussed in the next section.)

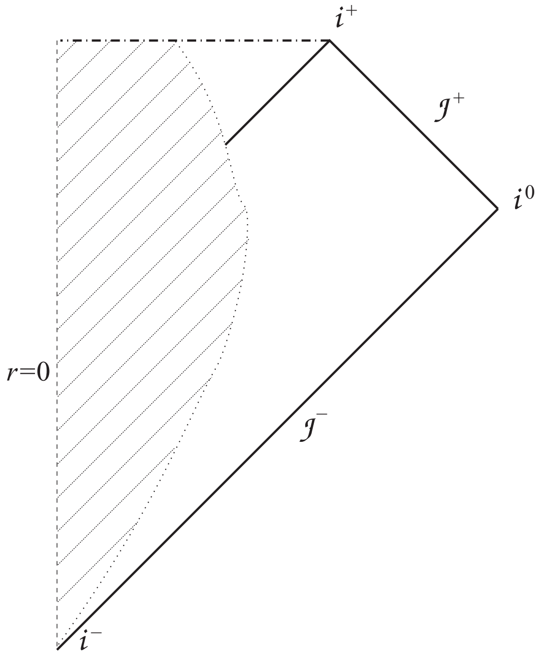

Fig. 5.19 Conformal diagram of Oppenheimer–Snyder gravitational collapse. The hashed region is the homogeneous piece representing the collapsing star, which cuts off most of the Kruskal extension of the vacuum Schwarzschild solution.

The Oppenheimer–Snyder matching can be used to construct the conformal diagram of this collapse process. We start with the conformal diagrams of FLRW and Schwarzschild space-times, and, as drawn in Fig. 5.18, cut out the parts used in the matching as interior or exterior space-times, respectively. In the closed model, for instance, we can start with a time when $\dot{a} = 0$, such that our collapse commences with a surface at rest. We choose a radius R satisfying $1 - kR^2 > 0$, to ensure that $R = \text{const}$ is in fact a surface in the interior of the cosmological space-time, and $aR > 2GM$ to have a star surface initially outside the horizon as seen from the exterior. To the future, we have $\dot{a} < 0$ and the surface collapses.

The interior will develop a singularity in the future, which must match with the singularity in the Schwarzschild geometry. This can only happen if the surface $r = R$ intersects the horizon, at a time when $2GM = \chi = Y = a(\tau)R$ is satisfied. At this time, we have $\dot{Y}^2 = (Y'/X)^2$ and thus $\dot{a}^2 = 1 - kR^2 > 0$, i.e. the shell is still collapsing. From the behavior of isotropic models, we know that, after $\dot{a} = 0$, $\dot{a}^2$ grows monotonically and diverges at a finite τ. The horizon condition $\dot{a}^2 = 1 - kR^2$ is thus satisfied exactly once, and the horizon is always crossed but never re-exited. The matched picture provides a conformal diagram of collapsing matter with a regular past, a horizon, a trapped region and a future singularity as in Fig. 5.19. Most of the Kruskal extension of the exterior Schwarzschild space-time is cut off and replaced by the matter interior.

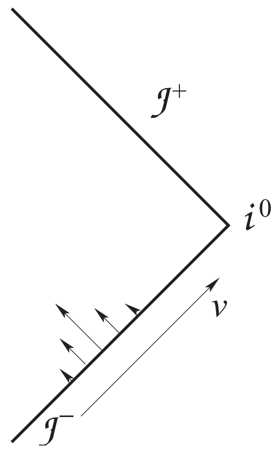

Fig. 5.20 Null dust falling in from $\mathcal{J}^-$. The energy flow is a function of the null coordinate v which varies along $\mathcal{J}^-$.

5.4.4 Vaidya solution

The Vaidya solution

$$\mathrm{d}s^2 = -\left(1 - \frac{2GM(v)}{r}\right)\mathrm{d}v^2 + 2\mathrm{d}v\mathrm{d}r + r^2\mathrm{d}\Omega^2\,, \tag{5.67}$$

found by Vaidya (1953), provides an asymptotically flat space-time that contains collapsing matter. It depends on one free function $M(v) \geq 0$, required to be non-decreasing. As a solution to Einstein's equation, it is sourced by an energy-momentum tensor

$$T_{ab} = \frac{G}{4\pi r^2}\frac{\mathrm{d}M}{\mathrm{d}v}(\nabla_a v)(\nabla_b v) \tag{5.68}$$

of null dust, corresponding to matter falling in along null lines; see Fig. 5.20: the asymptotic Killing vector $\xi^a = (\partial/\partial v)^a$ provides the matter flux $j_a = -T_{ab}\xi^b \propto -\nabla_a v$, and thus $j^a \propto -(\partial/\partial r)^a|_{v=\mathrm{const}}$. Global properties of the Vaidya space-time have been analyzed by Hiscock et al. (1982), Waugh and Lake (1986), Girotto and Saa (2004).

In order to model spherically symmetric collapsing matter, we assume a situation depicted in Fig. 5.21 in which $M(v) = 0$ for $v < 0$ and $M(v) = M$ constant for $v > v_0 > 0$. In between, matter is falling in along null lines from $\mathcal{J}^-$ such that $M(v)$ is increasing from zero to M between $0 < v < v_0$. We have a piece of Minkowski space-time for $v < 0$, a piece of the Schwarzschild solution for $v > v_0$ and part of Vaidya space-time in the region $0 < v < v_0$ where the collapse takes place. There are two matching surfaces along the null lines $v = 0$ and $v = v_0$. Matching conditions are easy to satisfy: The only discontinuity in the energy-momentum tensor is in the component T_{vv} at null lines where $\mathrm{d}M/\mathrm{d}v$ is not continuous. The main quantity required to be continuous is the expansion $\theta_\ell \propto 1 - 2GM(v)/r$, which is indeed continuous provided that $M(v)$ is continuous. Thus, the null-junction conditions are satisfied, and our matched space-time presents a solution to Einstein's equation everywhere.

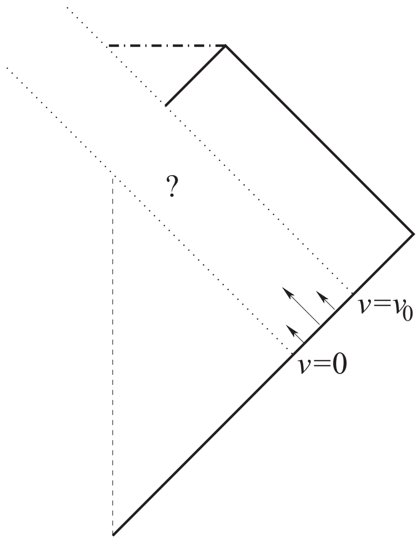

Fig. 5.21 Minkowski space in the past is matched to Schwarzschild space-time in the future through an intermediate region where null dust falls in during a Vaidya collapse phase.

We now model the collapse region by a linear increase $GM(v) = \mu v$ for $0 < v < v_0$ with constant μ. As exploited by Hiscock et al. (1982) for the construction of conformal diagrams as follows, this choice has the advantage that a conformal Killing vector field

$$\xi^a = v \left(\frac{\partial}{\partial v}\right)^a + r \left(\frac{\partial}{\partial r}\right)^a =: \left(\frac{\partial}{\partial \zeta}\right)^a \tag{5.69}$$

exists (Exercise 5.13), expressed here in new coordinates $z = v/r$, $\zeta = \log v$. In these coordinates, we have $z = \text{const}$ along the trajectories of the conformal Killing vector field: $\xi^a \nabla_a z = 0$. The range of the new coordinates is $0 < z < \infty$, $-\infty < \zeta < \log v_0$, and we obtain the line element

$$\begin{aligned} \mathrm{d}s^2 &= -\left(1 - \frac{2\mu v}{r}\right)\mathrm{d}v^2 + 2\mathrm{d}v\mathrm{d}r + r^2\mathrm{d}\Omega^2 \\ &= e^{2\zeta}\left(-\left(1 - 2\mu z - \frac{2}{z}\right)\mathrm{d}\zeta^2 - \frac{2}{z^2}\mathrm{d}\zeta\mathrm{d}z + \frac{1}{z^2}\mathrm{d}\Omega^2\right). \end{aligned} \tag{5.70}$$

We recognize the conformally symmetric nature because the ζ-dependence occurs only in the conformal factor. In what follows, we will mainly analyze $e^{-2\zeta}\mathrm{d}s^2$, which is sufficient for a conformal diagram except at places where $e^{-2\zeta} = 0$, at which the diagram will be extendable for the actual space-time (5.70).

As in Chapter 5.3, for the construction of conformal diagrams it is useful to introduce double-null coordinate, making the radial-time part of the metric purely off-diagonal. To do so, we eliminate the term $\mathrm{d}\zeta^2$ in the line element by introducing a new coordinate

$\eta := \zeta + f(z)$ such that

$$\left(1 - 2\mu z - \frac{2}{z}\right) f'(z) = \frac{2}{z^2} \,. \tag{5.71}$$

We then have

$$-\left(1 - 2\mu z - \frac{2}{z}\right) \mathrm{d}\zeta \mathrm{d}\eta = -\left(1 - 2\mu z - \frac{2}{z}\right) \mathrm{d}\zeta^2 - \frac{2}{z^2} \mathrm{d}\zeta \mathrm{d}z$$

as a representation of the line element in double-null form:

$$\mathrm{d}s^2 = e^{2\zeta} \left(-\left(1 - 2\mu z - \frac{2}{z}\right) \mathrm{d}\zeta \mathrm{d}\eta + \frac{1}{z^2} \mathrm{d}\Omega^2\right) \,.$$

The differential equation (5.71) is solved by the integral

$$f(z) = -2 \int^z \frac{\mathrm{d}x}{x(2\mu x^2 - x + 2)} \tag{5.72}$$

of a function with poles at $z = 0$ and $z_\pm = \frac{1}{4}\mu^{-1}(1 \pm \sqrt{1 - 16\mu})$. As a consequence, the new coordinate η is not always globally defined, since $f(z)$ may diverge at finite values of z, and we have to discuss different cases, depending on the behavior of the poles.

5.4.4.1 *Spacelike singularity*

For $\mu > 1/16$, there is only one real pole at $z = 0$ and we always have $2\mu z^2 - z + 2 > 0$. The function (5.72) is integrated to

$$f(z) = -\log z + \frac{1}{2} \log(2\mu z^2 - z + 2) - \frac{1}{\sqrt{16\mu - 1}} \arctan \frac{4\mu z - 1}{\sqrt{16\mu - 1}} \tag{5.73}$$

providing a coordinate $\eta = \zeta + f(z)$ that is well defined for all allowed values of z. In particular, we have $\eta = \zeta - \log z + g(z)$ with a function $g(z)$ bounded at $z = 0$.

The collapse region is described by a global double-null pair (ζ, η), in which we have the following limiting behaviors:

- $\eta \to \infty$ is reached for any finite ζ if $z \to 0$. For $z \to \infty$, $\eta \to \pm\infty$ diverges (the sign depending on the value of μ in the allowed range).
- $\zeta \to -\infty$ for finite η if $z \to 0$ in such a way that $ze^{-\zeta}$ remains finite, and $\zeta \to \infty$ for finite η if $z \to \infty$.

This asymptotic behavior of null coordinates determines the boundaries in the sense of $\mathcal{J}^\pm$, but there is also a singular part of the boundary: for the Vaidya metric, we have a curvature invariant of the form

$$R^{abcd} R_{abcd} = \frac{48 G M(v)^2}{r^6} = 48\mu^2 z^6 e^{-4\zeta} \tag{5.74}$$

which diverges for $z \to \infty$ when ζ remains finite.

In order to disentangle the different parts of the boundary, we first look at the case where $v_0 \to \infty$, i.e. the mass function $M(v)$ increases linearly from zero at $v = 0$ to infinity

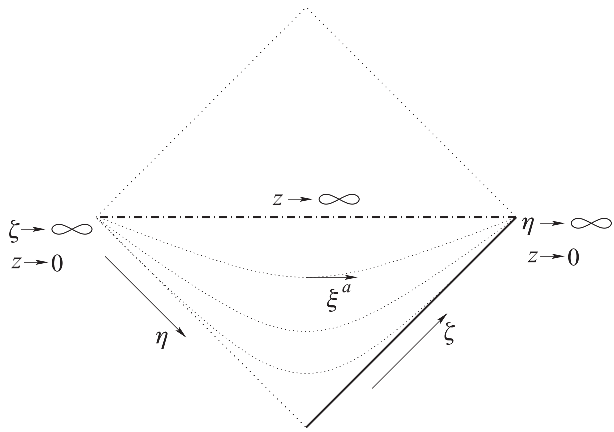

Fig. 5.22 The Vaidya region for $\mu > 1/16$. The solid boundary is $\mathcal{J}^-$, the bottom dotted one is extendable, and the dash-dotted one is singular. The upper triangle is cut off by the singularity.

without ever stopping at a constant value. We obtain a space-time region of triangular shape, with boundaries of three different types; see Fig. 5.22. For $\zeta = \log v$, we use the full real range in this case, but space-time is extendable through $\zeta \to -\infty$ where the conformal factor $e^{2\zeta}$ vanishes: the actual metric (5.70) serves as a conformal completion of the ζ-independent line element that we are analyzing. In fact, in the original coordinate v, the region where $\zeta \to -\infty$ corresponds to a finite boundary. The boundary is an extended region if $\eta = \zeta + f(z)$ is allowed to run through a set of finite values, which is possible provided that $z = 0$ there: only then can the divergence of ζ be cancelled by a pole in $f(z)$. As seen above, the approach at finite η always happens in such a way that $ze^{-\zeta}$ remains finite, such that the curvature invariant behaves like $e^{2\zeta}$ and vanishes; this part of the boundary is thus non-singular. The allowed range of η provides the first edge of our triangular boundary.

The other two edges correspond to physical, non-extendable boundaries of space-time. Next, we take $\eta \to \infty$ which allows all values of ζ if $z \to 0$. Also here, curvature remains finite and presents a regular part of the boundary. Now, the conformal factor $e^{2\zeta}$ is finite; thus, the boundary is asymptotic and presents past null infinity, since it maps out all finite values of $v > 0$ in our allowed range. The final part of the boundary is obtained for finite ζ and η, but $z \to \infty$ (and thus $r \to 0$). This provides a singular boundary because the curvature invariant diverges. It is not a null boundary because both null coordinates are finite, which is possible at $z \to \infty$ when the logarithmic singularities in $f(z) = \eta - \zeta$ cancel for $z \to \infty$ (and ζ is independent of z). This completes the conformal picture, Fig. 5.22, of the diagram in the region of $0 < v < \infty$, provided that the linear increase of $M(v)$ goes on forever.

However, we are only interested in a region of finite range $0 < v < v_0$ after which the infall of matter stops and we settle to a black hole of mass $GM = GM(v_0) = \mu v_0$. We take a finite part of the conformal diagram obtained by cutting off an upper-right triangle

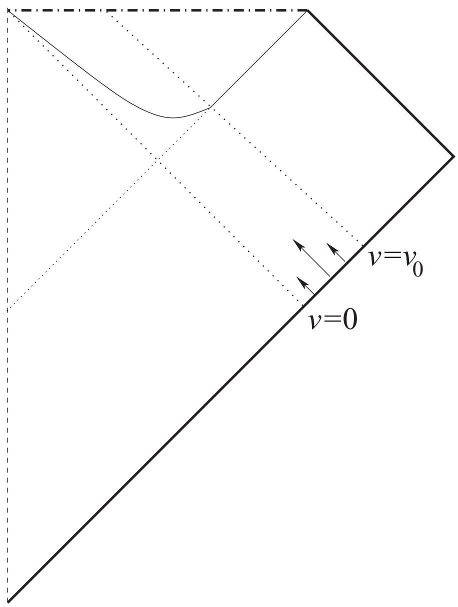

Fig. 5.23 Conformal diagram of Vaidya collapse for $\mu > 1/16$, matched to Minkowski space in the past and Schwarzschild space-time in the future.

corresponding to $v > v_0$. Above the open edge, we match to Schwarzschild space-time with mass M, connecting the two $\mathcal{J}^-$ and the singular regions and adding in the $\mathcal{J}^+$, i^+ and i^0 of Schwarzschild. To the lower left, we extend the diagram by matching to Minkowski space because $M(0) = 0$ is approached there. This provides the complete conformal diagram, Fig. 5.23, of a collapse model in which infalling matter is described by null dust. All boundary points in this diagram are either asymptotic, singular or symmetry centers; the diagram is thus complete. (Again, the matching is possible because the relevant parameters required to be continuous only depend on $M(v)$, not $M'(v)$.)

Qualitatively, the diagram is the same as that obtained in the Oppenheimer–Snyder model. But there is a new feature: in the Vaidya region, we can past-extend the horizon of the Schwarzschild region by the boundary of the region of spherical trapped surfaces, described by $r = 2M(v)$ (as in Example 5.3). Since $M(v)$ is not constant but decreases to the past, this line recedes to the interior and touches the singularity at $v \to 0$. Such a horizon behavior is very different from that of the event horizon, which is obtained by past-extending the Schwarzschild horizon as a null surface. In particular, while the region of trapped surfaces cannot overlap with the Minkowski patch, the event horizon does enter the initially empty vacuum region even though no causal contact with the infalling null matter could have existed at these times (and no trapped surfaces have formed). This illustrates the global nature of the event horizon, which requires knowledge of the entire future to determine its complete extension.

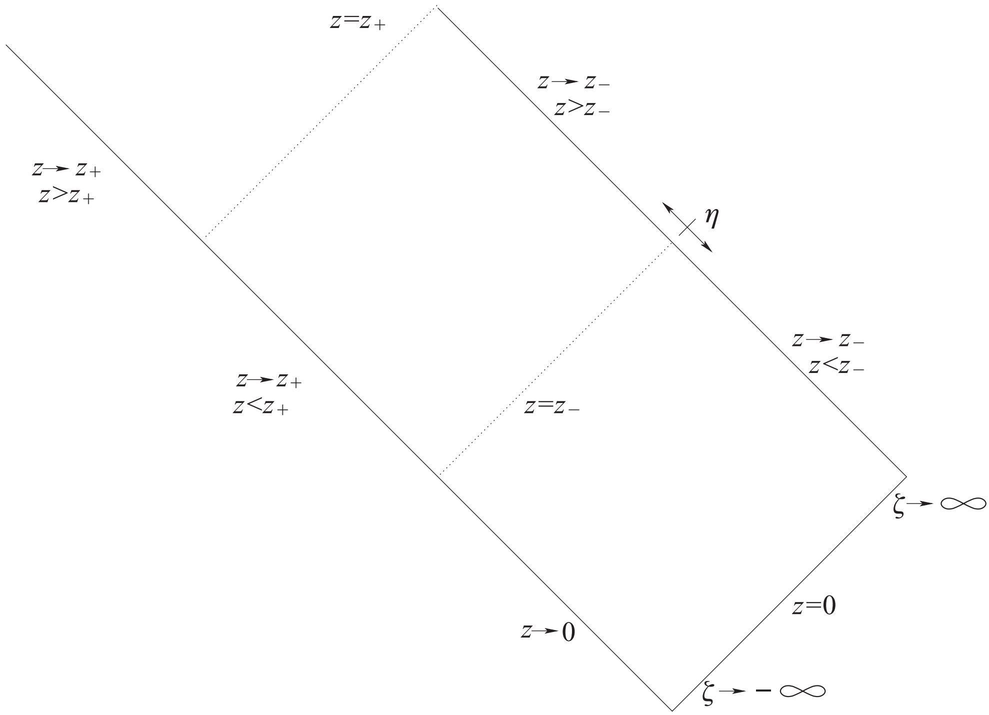

Fig. 5.24 Coordinates in the Vaidya region for $\mu < 1/16$.

5.4.4.2 Null singularity

For $0 < \mu < 1/16$, we integrate (5.72) to obtain

$$\eta = \zeta - \log z - \frac{2\mu^2 z_-}{\sqrt{1-16\mu}} \log|z - z_+| + \frac{2\mu^2 z_+}{\sqrt{1-16\mu}} \log|z - z_-| \tag{5.75}$$

which is ill defined at $z_\pm = \frac{1}{4}\mu^{-1}(1 \pm \sqrt{1-16\mu})$. We thus require three different patches of double-null coordinates, corresponding to $0 < z < z_-$, $z_- < z < z_+$ and $z > z_+$.

Also here, the collapse region is bounded by two null lines parameterized by ζ and η. A diverging η will again correspond to some part of $\mathcal{J}^-$, while the boundary lines spanned by all finite η provide null matching surfaces to Minkowski space and the Schwarzschild solution, respectively. As before, we first allow v to grow without bound, and will later cut out the relevant part up to $v = v_0$. The range for ζ is thus the whole real line.

At $\zeta \to \infty$, we must consider two ranges of η, drawn in Fig. 5.24, since η diverges at $z \to 0$ and at $z \to z_+$. For $z \to z_-$, on the other hand, η may remain finite, depending on the precise approach of $\zeta \to \infty$ so as to cancel the ζ-divergence with that of the last logarithm in (5.75). The null line $\zeta \to \infty$ thus has two components, both of which correspond to points where $z \to z_-$ with finite η. On one component, z_- is approached from below and on

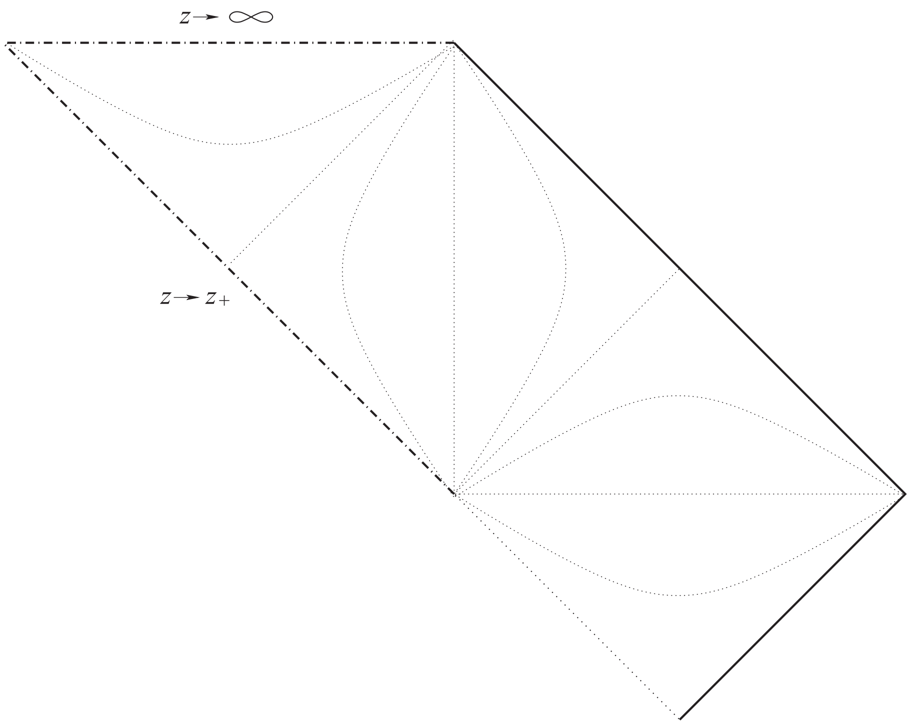

Fig. 5.25 Conformal diagram of a Vaidya region for $\mu < 1/16$ with unbounded linear increase of the mass. Coordinate lines of constant z are dotted. Dash-dotted lines are singularities.

the other from above. Also, $z \to 0$ and $z \to z_+$ lie on the boundary part with $\zeta \to \infty$, but since they require $\eta \to \infty$, too, they only present single points in the conformal diagram.

For $\zeta \to -\infty$, on the other hand, there are three components: we necessarily have $\eta \to -\infty$ for $z \to z_-$, giving a single point on the boundary, but η can be finite if $z \to 0$ or $z \to z_+$. There are now three boundary contributions, corresponding to $z \to 0$ and $z \to z_+$ from below as well as from above.

Not all the boundary contributions are regular (see Fig. 5.25): curvature diverges whenever $z \to \infty$ and ζ remains finite, or when $\zeta \to -\infty$ and $z \neq 0$. The first possibility provides a space-like singularity, as in the previous case, but the second possibility is a null singularity because it lies at constant values of the null coordinate $\zeta \to -\infty$. It is an extended region rather than a single point because it encompasses the whole boundary where $z \to z_+$. On the remaining $\zeta \to -\infty$ part, where $z \to 0$, on the other hand, curvature is finite but the conformal factor $e^{2\zeta}$ of the metric (5.70) becomes zero. Space-time is extendable across this boundary part where $v = 0$, and here we can match to a past Minkowski region.

To match to Schwarzschild in the future, we take the collapse region only up to a finite value of ζ corresponding to $v = v_0$. Here, the mass has grown to $GM = \mu v_0$, which we

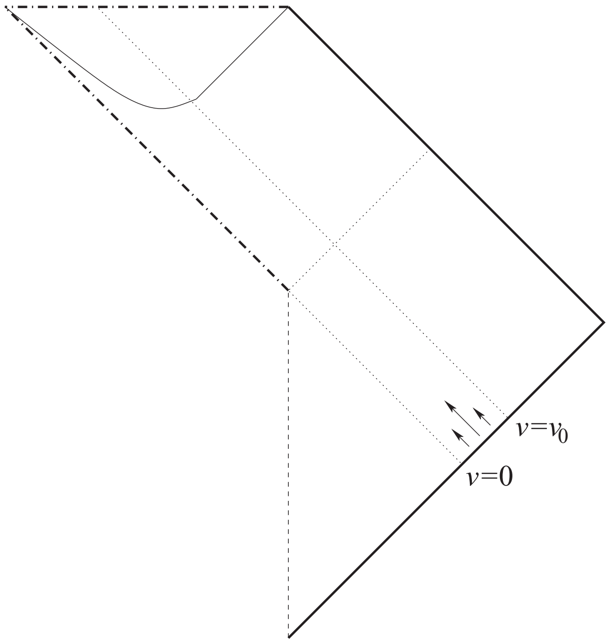

Fig. 5.26 Conformal diagram of Vaidya collapse for $\mu > 1/16$, matched to Minkowski space in the past and Schwarzschild space-time in the future. A naked null singularity and a Cauchy horizon develop.

use for the Schwarzschild region to be matched. The Schwarzschild region then provides $\mathcal{J}^+$ of the complete diagram, Fig. 5.26, and the part of the singularity that matches the space-like singularity of the collapse region. It also shows a horizon, which is located at $z_{\rm hor} = v_0/r_{\rm hor} = 1/2\mu > z_+$. Due to $z_{\rm hor} > z_+$, there is always some part of the null singularity that lies outside the event horizon: the singularity is *naked* because it can emit signals to $\mathcal{J}^+$. The boundary of the causal future of the naked singularity is called a *Cauchy horizon*. Evolution from an initial spatial slice is well posed only outside the Cauchy horizon. (In general, there is a locally naked and null singularity if $\dot{M}(0) < 1/16$, even if $M(v)$ is not linear.)

5.4.5 Cosmic-censorship conjecture

Space-times with naked singularities, as encountered in some of the Vaidya solutions, are not globally hyperbolic and fail to provide predictivity in the sense of an initial-value formulation. This physically unpleasant situation has given rise to the *cosmic-censorship conjecture* which posits that gravitational collapse of *generic* initial data always leads to singularities that are covered behind horizons. The conjecture has not been proven in general, but, so far, no counterexamples have been provided.

Here, it is important to keep in mind the requirement of generic initial data. As we have seen in the discussion of Vaidya solutions, large classes of models do exist in which naked singularities form in gravitational collapse. However, due to their symmetry, these solutions do not provide generic initial conditions. As analyzed by Podolský and Svítek (2005), one can perturb around the Vaidya solutions, for instance in the form of *Robinson–Trautman metrics*

$$\mathrm{d}s^2 = -\left(K + 2r\partial_v \log f - 2\frac{M}{r}\right)\mathrm{d}v^2 + 2\mathrm{d}v\mathrm{d}r + \frac{r^2}{f^2}\mathrm{d}\Omega^2$$

with the mass function $M(v)$, as in Vaidya space-times, but also including an arbitrary perturbation by a non-spherical $f(v, \vartheta, \varphi)$. The perturbation f, for this class of models, then defines $K = f^2(1 + \Delta_2 \log f)$ where Δ_2 is the Laplacian on the unit 2-sphere with line element $\mathrm{d}\Omega^2$.

With the stress-energy tensor of Vaidya, this is a solution to Einstein's equation if

$$\frac{\partial f}{\partial v} = \frac{f\Delta K}{12M(v)}\,.$$

A collapse region perturbed in this way can be matched to past Minkowski and future Schwarzschild as before, but now it turns out that solutions for f diverge for $v \to \infty$. Thus, the complete Vaidya solution is not stable against perturbations, and the naked singularities they show are not generic. This instability of a physical phenomenon, in this case the formation of naked singularities, means that the phenomenon can have only limited meaning. (An example of a stable phenomenon is the general occurrence of singularities of the general type in terms of geodesic incompleteness, which is a generic feature according to the singularity theorems.)

Incidentally, Robinson–Trautman metrics also exhibit an interesting behavior regarding their differentiability. Just like Vaidya solutions, they can be matched to Minkowski space at $v = 0$, where f cannot be smooth. Instead, for $M(v) = \mu v$ we have the following two cases:

1. If $2/\mu$ is an integer, f is $30/\mu - 1$ times differentiable at the matching surface. Since $\mu \leq 2$ for $2/\mu$ to be integer, this means that f is at least 14 times differentiable. In the limiting case of $\mu = 1/16$, it is 479 times differentiable, but not more.
2. Otherwise, f is $\{2/\mu\}$ times differentiable, which designates the largest integer smaller than $2/\mu$. If $\mu > 2$, f is not differentiable but continuous.

5.5 Horizons

A horizon, in contrast to a sharp surface such as the boundary of a star, captures the notion of the border of a black hole. We have already seen several examples in explicit solutions, as well as the importance of horizons to characterize the occurrence of strong gravitational forces able to capture even light. While the physical picture as a region trapping light is intuitive, the general mathematical situation is not so clear. There are different definitions of horizons, which in many cases lead to different pictures of the behavior. A precise definition, however, would be important, for instance to discuss the merger process of two black holes; if we are not able to locate their horizons precisely, we cannot tell when a single horizon has formed, indicating that the black holes have merged. Some of the properties associated with different notions of horizons are discussed in this section.

5.5.1 Notions of horizons

The *event horizon*, defined as the boundary of the past of $\mathcal{J}^+$, is a global concept. The past is the set of all points that can be connected to $\mathcal{J}^+$ by future-pointing timelike or null curves. Thus, the event horizon is always a null submanifold. It requires global knowledge of space-time due to the use of $\mathcal{J}^+$ in its definition.

A more local definition of horizons uses *future-trapped surfaces* which can be analyzed without reference to asymptotic regions: as defined in Chapter 5.1.2, these are closed spatial 2-surfaces S such that both future-pointing null normal congruences emanating from them have negative expansion. A *marginal future-trapped surface* is the limiting case of a closed spatial 2-surface such that the expansion of one future-pointing null normal congruence vanishes and the other one is negative. Future-trapped surfaces can only lie within an event horizon in asymptotically flat situations, because the necessary presence of conjugate points along both null congruences prevents null curves from reaching $\mathcal{J}^+$; thus, they cannot be in the past of $\mathcal{J}^+$. The motivation for the following definitions is that the limiting case of marginal trapped surfaces may provide a good local definition of a horizon, which in fact in some cases agrees with the event horizon.

The definition of an *apparent horizon* is based on a foliation of an asymptotically flat space-time into spatial slices Σ. On each slice, the set of outermost marginal trapped surfaces lying in Σ defines the place where the apparent horizon intersects the slice. (Outermost surfaces are here defined as those closest to the asymptotic region along the slice.) Stacking up all slices of the foliation and the outermost marginal trapped surfaces within them then provides the 3-dimensional manifold of the apparent horizon in space-time. While this notion is more local than that of an event horizon, it has the disadvantage of being foliation-dependent. Moreover, for a given foliation, only trapped surfaces within an entire Σ are considered, a procedure that may overlook other trapped surfaces (as discussed explicitly by Wald and Iyer (1991)). It turns out that the apparent horizon may be discontinuous, for instance in the collapse of two black holes. This concept of a horizon, as described by Thornburg (1996, 2007) is used mostly in numerical simulations, which are already based on the evolution of data on spatial slices and provide an underlying foliation by their setup.

A *trapping horizon* as defined by Hayward (1994) is a 3-dimensional surface H in space-time which allows a foliation by marginal future-trapped surfaces. This concept requires neither asymptotic structures nor a foliation of space-time and thus eliminates the disadvantages of event and apparent horizons. (The notion of a trapping horizon has

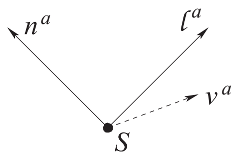

Fig. 5.27 Two null vectors n^a and ℓ^a, and the tangent v^a to the horizon, with cross-section by the 2-sphere S.

been used to introduce various types of horizon based on the behavior of marginally trapped surfaces, most of which are situated in the class of isolated and dynamical horizons reviewed by Ashtekar and Krishnan (2004), and with an application to numerical relativity by Dreyer et al. (2003) and Schnetter et al. (2006).) Given a trapping horizon, we have 2-dimensional cross-sections S that are marginal future-trapped. For each cross-section, there are two future pointing null normals ℓ^a and n^a, cross-normalized to $\ell^a n_a = -1$, one of which, say ℓ^a, has vanishing expansion, $\theta_\ell = 0$. On the cross-section S, we have an induced metric $q_{ab} = g_{ab} + 2n_{(a}\ell_{b)}$.

5.5.2 Spacelike properties

For a marginal future-trapped surface, we have by definition $\theta_\ell = 0$, $\theta_n < 0$. We call the horizon an *outer trapping horizon* if $n^a\nabla_a\theta_\ell < 0$ at any of its 2-sphere cross-sections S, where we extend ℓ^a to a neighborhood of S. In this case, the trapped region lies in the direction of n^a. The null version of the Raychaudhuri equation

$$\ell^a\nabla_a\theta_\ell = -\frac{1}{2}\theta_\ell^2 - \sigma^\ell_{ab}\sigma_\ell^{ab} - R_{ab}\ell^a\ell^b$$

with $\omega_{ab} = 0$, due to hypersurface orthogonality to S, then implies properties of the trapping horizon as a hypersurface in space-time.

By definition, S is a spacelike submanifold of the trapping horizon H. The normal to S within H is proportional to $v^a = \ell^a + \alpha n^a$, as in Fig. 5.27, since ℓ^a and n^a generate the normal space to S. Along H, $\theta_\ell = 0$ stays constant and we have

$$0 = v^a\nabla_a\theta_\ell = \ell^a\nabla_a\theta_\ell + \alpha n^a\nabla_a\theta_\ell\,;$$

thus,

$$\alpha = -\frac{\ell^a\nabla_a\theta_\ell}{n^b\nabla_b\theta_\ell}\,.$$

For an outer trapping horizon, $n^a\nabla_a\theta_\ell < 0$, while $\ell^a\nabla_a\theta_\ell \leq 0$ by the Raychaudhuri equation if the null-energy condition, and consequently $R_{ab}\ell^a\ell^b \geq 0$, is satisfied. Thus, the normal to S in H has norm $g_{ab}v^av^b = -2\alpha \geq 0$ and is spacelike. Since the tangent space to H is generated by v^a together with the tangent space to S, all vectors tangent to H are spacelike: future outer trapping horizons are generically spacelike if the null-energy condition holds. They are null if their normal congruence ℓ^a is shear-free and $R_{ab} = 0$ on them. Only in this case, which can be interpreted as the absence of gravitational radiation and matter at the horizon, can the trapping horizon agree with the event horizon.

Example 5.6 (Trapping horizon in the Vaidya space-time)
The Vaidya space-time is spherically symmetric, making it easy to find spherical marginally trapped surfaces. Using only spherical surfaces, however, is a restriction of the general set of trapped surfaces to be considered according to the definition of a trapping horizon; we will only obtain a lower bound for the radius of the trapping horizon in this way: there may well be non-spherical trapped surfaces outside the region of spherical trapped surfaces. Such trapped surfaces have indeed been found numerically by Krishnan and Schnetter (2006) as well as analytically by Ben-Dov (2007) in the Vaidya space-time; Eardley (1998) has conjectured that every point within the event horizon allows a trapped surface through it, and Bengtsson and Senovilla (2009) have constructed trapped surfaces in the matched Vaidya space-time as constructed here even through points in the flat Minkowski region. The region of spherical marginal trapped surfaces satisfies $r = 2GM(v)$*, a surface which has the co-normal* $n_a = \nabla_a r - 2G(\nabla_a v)\mathrm{d}M/\mathrm{d}v$*. For collapsing matter, we have* $n^a n_a = -4G\mathrm{d}M/\mathrm{d}v < 0$ *as the timelike normal, and thus the surface* $r = 2GM(v)$ *is indeed spacelike as expected for a trapping horizon.*

Even though spherical trapped surfaces may not give the exact horizon position, the trapping horizon cannot agree exactly with the event horizon. The event horizon is always null, while the trapping horizon is null only if the shear of ℓ^a *as well as* $R_{ab}\ell^a\ell^b$ *vanish along it. For the Vaidya space-time, we have a stress-energy tensor* $T_{ab} = (4\pi r^2)^{-1}(\mathrm{d}M/\mathrm{d}v)n_a n_b$ *where* $n_a = \nabla_a v$ *is the ingoing null normal to the marginal trapped surfaces slicing the horizon. Using cross-normalization with the outgoing null normal* ℓ^a*, we have* $R_{ab}\ell^a\ell^b = 8\pi G T_{ab}\ell^a\ell^b = 2Gr^{-2}\mathrm{d}M/\mathrm{d}v > 0$*, which does not vanish in the collapse region. Thus, in any region where the mass function strictly increases, the trapping horizon of a Vaidya space-time is spacelike and cannot agree with the event horizon. If there are different phases of matter infall, as in Fig. 5.28, even a null part of a trapping horizon can be distinct from a part of the event horizon.*

Properties of spacelike horizons can often be derived from the constraint equations, which must hold on any spacelike surface. In particular, according to Ashtekar and Krishnan (2003), balance laws follow from integrations of the constraints over regions of the horizon.

Example 5.7 (Balance of quasilocal angular momentum)
If we denote the unit normal to a spacelike horizon by N^a *(to distinguish it from the cross-normalized null normal* n^a *used conventionally in the context of horizons), the diffeomorphism constraint (3.64) with a matter contribution is*

$$C^a = -2\nabla_b p^{ab} + \sqrt{\det h} N_b T^{ab} = \sqrt{\det h}\left(-\frac{1}{8\pi G} D_b(K^{ab} - K^c{}_c h^{ab}) + N_b T^{ab}\right). \tag{5.76}$$

We now pick a foliation of the horizon into spherical cross-sections, choose a vector field ϕ^a *tangent to the cross-sections, and integrate the* $\phi^a C_a$ *over a finite part* H *of the horizon*

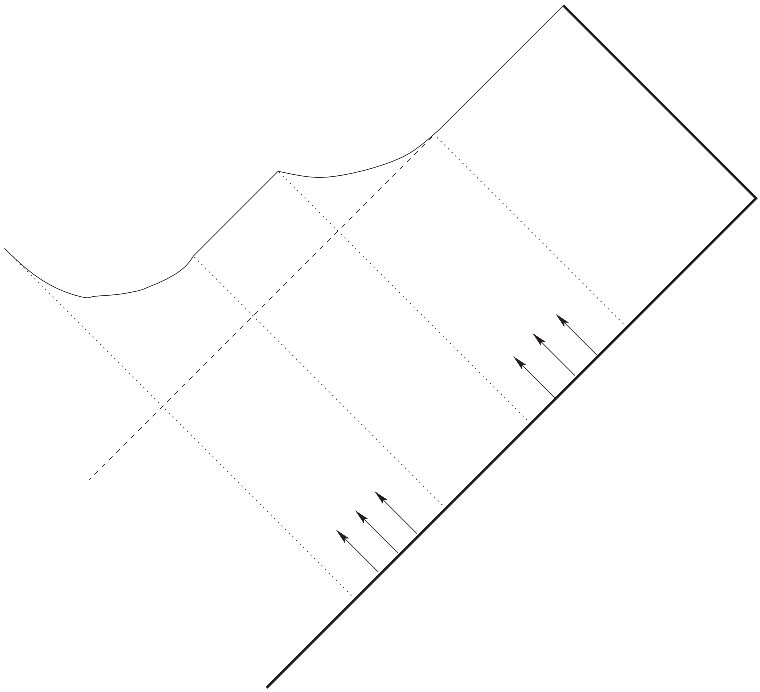

Fig. 5.28 Trapping horizon (solid thin) compared with the event horizon (dashed) in a collapse situation. Infalling matter is indicated by the arrows, realized in two regions separated from one another by a vacuum patch. The trapping horizon is null in vacuum regions, but may not fall on the event horizon even in this case.

bounded by two cross-sections S_1 and S_2. Integrating by parts, we have

$$\frac{1}{8\pi G}\left(\int_{S_2}\phi^a r^b K_{ab}\sqrt{\det q}\,\mathrm{d}^2y - \int_{S_1}\phi^a r^b K_{ab}\sqrt{\det q}\,\mathrm{d}^2y\right)$$
$$= \int_H \left(\phi_a N_b T^{ab} + \frac{1}{16\pi G}(K^{ab} - K^c{}_c h^{ab})\mathcal{L}_\phi q_{ab}\right)\sqrt{\det h}\,\mathrm{d}^3x \qquad (5.77)$$

with r^a normal to the cross-sections within H and the induced metric $q_{ab} = h_{ab} - r_a r_b$ on cross-sections. On the left-hand side, we recognize the change of Brown–York quasilocal (angular) momentum (3.73) between two cross-sections (for which the subtraction cancels out if it corresponds to a stationary metric). If ϕ^a is a Killing vector field of the cross-section metrics, the change of quasilocal (angular) momentum is given by the energy flow $\phi_a N_b T^{ab}$ through the horizon.

Ashtekar and Krishnan (2003) derive another balance law for the horizon area from the Hamiltonian constraint, which is in line with a large class of results about horizon area increase.

5.5.3 Area increase

There are several general properties, provable by different methods for trapping as well as event horizons, even though these concepts in general do not agree in terms of the submanifold they define. One of the main results states that the cross-section area of a horizon cannot decrease if energy conditions are satisfied. The area of cross-sections of a trapping horizon is determined by the induced metric $q_{ab} = g_{ab} + 2n_{(a}\ell_{b)}$, as $A = \int_{S^2} \mathrm{d}y^2 \sqrt{\det q}$. Its change along the horizon, derived by Hayward (1994), is given by its Lie derivative along the tangent vector $v^a = \ell^a + \alpha n^a$:

$$\begin{aligned} \mathcal{L}_v \det q &= \det q\, q^{ab} \mathcal{L}_v q_{ab} = \det q\, q^{ab} (\mathcal{L}_\ell q_{ab} + \alpha \mathcal{L}_n q_{ab}) \\ &= \det q\, (\theta_\ell + \alpha \theta_n) \end{aligned} \tag{5.78}$$

where we used

$$q^{ab} \mathcal{L}_\ell q_{ab} = q^{ab} \nabla_a \ell_b + q^{ab} (\ell_a \mathcal{L}_\ell n_b + n_a \mathcal{L}_\ell \ell_b) = \theta_\ell\,.$$

On the trapping horizon, $\theta_\ell = 0$ and $\theta_n < 0$. Moreover, we have seen that $\alpha < 0$ for a spacelike horizon, $\alpha = 0$ for a null horizon, and $\alpha > 0$ for a timelike horizon if that were allowed. Thus, the area must increase along a spacelike horizon, remains constant along a null horizon and decreases along a timelike horizon. This is consistent with the understanding that the black hole grows if stress-energy is falling through its horizon, which could either come from matter as $T_{ab}\ell^a \ell^b$ or from gravitational radiation as $\sigma^\ell_{ab} \sigma_\ell^{ab}$. Any contribution of these types, both positive provided that the null energy condition holds, results in a spacelike trapping horizon whose area grows. Again using the Vaidya space-time as an example, we can directly confirm this general observation at the boundary of spherical trapped surfaces whose radius is $r = 2M(v)$. The cross-sections at constant v thus have areas $A(v) = 4\pi r(v)^2 = 16\pi M(v)^2$, a function which increases if matter falls through.

A similar result holds for the event horizon as derived by Hawking (1972), although it is slightly different: the event horizon is always null, but still the cross-section area may decrease if matter is falling through. On the other hand, the area can never decrease if energy conditions hold.

5.5.4 Stationary horizons and black-hole thermodynamics

A special class of black holes is given by stationary ones, in which we have a null trapping horizon (non-growing in a stationary situation) such that there is a Killing vector field χ^a in a neighborhood that agrees with $\chi^a = \ell^a$ on the horizon. For instance, the Schwarzschild horizon is stationary with $\chi^a = (\partial/\partial t)^a$. In this case, the Killing vector field agreeing with ℓ^a on the horizon is the stationary Killing vector field of space-time. But this need not be the case in general, especially for rotating black holes. In the latter case, we have instead $\chi^a = \xi^a + \Omega \psi^a$ where ψ^a is a spacelike Killing vector field (for a rotational symmetry) and Ω the angular velocity.

In the case of a stationary horizon, the congruence defined by ℓ^a has special properties: by definition, it is expansion free (outward normal to marginal trapped surfaces) and rotation free (hypersurface orthogonal). Moreover, it is shear free because only this allows the trapping horizon to be null. In fact, ℓ^a defines a null geodesic congruence.

Using the horizon as a null surface in a space-time solution without a surface layer of energy-momentum, the matching analysis of Chapter 5.4.2 shows that $\omega = n_a\ell^c\nabla_c\ell^a$ is continuous across the horizon. (Note that the roles of ℓ^a and n^a, using conventional notations, are exchanged compared to our notation in the discussion of null-junction conditions.) If ω were discontinuous, spatial components of the stress-energy tensor were singular at the surface, with a singular part with coefficient $8\pi GT_{ab}^{\rm sing} \sim G_{ab}^{\rm sing} = -\Delta\omega q_{ab}$. If we wanted to place the interior of the stationary horizon in empty flat space, rather than the black-hole exterior, the matching could be consistent only if a surface pressure P given by the singular part of T_{ab} stabilized the horizon. This surface pressure, $8\pi GP = \Delta\omega$, implies that we have to exert a force on the whole surface to replace the pull of gravity normally active when the interior belongs to a black hole rather than a surface layer of sress-energy in Minkowski space. We identify the required counteracting pressure $\kappa = -\frac{1}{2}\Delta\pi GP = -4\omega$ as the *surface gravity* acting on the boundary of the interior region:

$$\kappa := -\frac{1}{2}n^a\ell^b\nabla_b\ell_a\,. \tag{5.79}$$

Since only derivatives along the stationary horizon are involved in κ, we can replace ℓ^a with the Killing vector χ^a, and use the Killing-vector property to write

$$\kappa = -\frac{1}{2}n^a\chi^b\nabla_b\chi_a|_{\rm Hor} = \frac{1}{2}n^a\chi^b\nabla_a\chi_b|_{\rm Hor} = \frac{1}{4}n^a\nabla_a(\chi^b\chi_b)|_{\rm Hor}\,.$$

(Notice that the last derivative is not along the horizon where $\chi^b\chi_b = \ell^b\ell_b = 0$ would be constant.) From this equation, we can see that κ is constant along ℓ^a: using $\nabla^a(\chi^b\chi_b) = -4\kappa\chi^a$ from the definition of κ and with cross-normalization $\ell^a n_a = -1$ (which on the horizon holds for χ^a instead of ℓ^a, too), we have on the horizon

$$\begin{aligned}-4\ell^a\mathcal{L}_\ell\kappa &= \mathcal{L}_\ell(-4\kappa\chi^a) = \mathcal{L}_\chi\nabla^a(\chi^b\chi_b) = \chi^c\nabla_c(\nabla^a(\chi^b\chi_b)) - \nabla^c(\chi^b\chi_b)\nabla_c\chi^a\\ &= \nabla^a(\chi^c\nabla_c(\chi^b\chi_b)) - (\nabla^a\chi^c)\nabla_c(\chi^b\chi_b) - (\nabla_c\chi^a)\nabla^c(\chi^b\chi_b) = 0\end{aligned}$$

where the first term vanishes because it is a derivative of $\chi^b\chi_b$ along the horizon where it is constant, and the last terms cancel each other because χ^a is a Killing vector field. For stationary black holes, the gravitational force thus remains constant.

Moreover, if the dominant energy condition holds, surface gravity is constant all over the spatial cross-sections of the horizon, even in situations that are not spherically symmetric. In fact, one can show that $\chi_{[d}\nabla_{c]}\kappa = -\chi_{[d}R_{c]}^f\chi_f = 0$ which vanishes because $T^{cf}\chi_f$, given the dominant energy condition, must be future-directed timelike or null. Due to $T_{ab}\chi^a\chi^b = 0$ along the stationary horizon, this is possible only if $T^{cf}\chi_f \propto \chi^c$, and thus $R^{cf}\chi_f \propto \chi^c$, making $\chi_{[d}\nabla_{c]}\kappa = -\chi_{[d}\chi_{c]} = 0$. Contraction with spatial tangent vectors orthogonal to χ^a then shows that the surface gravity is the same everywhere on the stationary horizon.

Surface gravity plays a role in parameter changes between different black holes, which one may think of as adiabatic processes slowly changing the stationary horizon. In such processes, one can show that mass, surface area and angular momentum can change only in ways respecting the relation

$$\delta M = \frac{1}{8\pi G}\kappa\delta A + \Omega\delta J \,. \tag{5.80}$$

This relationship is very similar to thermodynamics. In fact, by analogy Bardeen et al. (1973) formulate several laws of *black-hole thermodynamics*:

0th law: Surface gravity κ is constant on stationary horizons and plays a role analogous to temperature.

1st law: In adiabatic processes, the mass changes by (5.80) where $\kappa\delta A/8\pi G$ is analogous to $T\delta S$, in agreement with an identification of surface gravity as temperature. This suggests that the surface area is to be considered as a measure analogous to entropy.

2nd law: The surface area does not decrease provided that positive energy holds (which corresponds to stability). This is again consistent because we have interpreted the area as being analogous to entropy.

This set of analogies suggests that surface gravity κ can be interpreted as a measure for the temperature of the black hole, and A as its entropy. Then, the term $\kappa\delta A$ in (5.80) plays the role of $T\delta S$ in thermodynamics. These laws are formal analogies,[2] but they can, quite surprisingly, be substantiated: no-hair theorems show that stationary black-hole solutions are determined only by three parameters, their mass together with angular momentum and electric charge. (Black-hole uniqueness theorems are reviewed by Heusler (1998).) Thus, only a few macroscopic parameters determine the whole black hole as a large system. Moreover, detailed calculations started by Bekenstein (1973) show that the black-hole area increases if matter is dropped in adiabatically by precisely the amount required to make $S_{\rm total} = S_{\rm matter} + S_{\rm BH}$ non-decreasing. If the black-hole area would be ignored, on the other hand, entropy would decrease if matter is dropped into the black hole where it disappears from outside view.

There is, however, a problem with this picture: black holes would have negative heat capacity, defined as the derivative $\partial E/\partial T$ keeping the other intensive variables constant. For the Schwarzschild solution, for instance, we have $\kappa = 1/4M$, and thus $E \propto 1/T$ with $\partial E/\partial T < 0$. A system with negative heat capacity should radiate thermally, but no radiation can escape from a black hole. There is also a dimensional problem: entropy is dimensionless, while area has the dimension length squared. From the fundamental constants involved in classical general relativity, one cannot form a quantity of dimension length to match the dimensions of entropy and area.

One indication for a resolution comes from the fact that black holes could radiate if stress-energy components were to become negative. This is usually not allowed for classical matter,

[2] Moreover, they can be derived in this form for the different versions of horizons that have been discussed, which in general do not correspond to the same physical surface. Thus, it seems difficult to attribute too much physical meaning to the laws of black-hole thermodynamics. Nevertheless, they still stimulate many questions about classical and quantum gravity.

but positive energy conditions can be violated by quantum fields. For instance, uncertainty relations can provide negative energy for brief times due to fluctuations. More precisely, there are general quantum inequalities for quantum fields on curved space-times as derived, for instance, by Pfennig and Ford (1998), which take the form

$$\int \rho(t)g(t)\mathrm{d}t \geq -\frac{c}{\Delta T^4} \tag{5.81}$$

where $g(t)$ is a sampling function of width ΔT and c is a constant. This inequality does provide a lower bound on the energy contained in a space-time region, but this lower bound is negative. That quantum physics can provide negative energies can also be seen based on superpositions: given a system with a ground state $|0\rangle$ of vanishing energy, $\langle 0|E|0\rangle = 0$, and an excited state $|1\rangle$, the energy in the superposition

$$|\psi\rangle := \frac{1}{\sqrt{1+\epsilon^2}}(|0\rangle + \epsilon|1\rangle)$$

is

$$\langle\psi|E|\psi\rangle = \frac{1}{1+\epsilon^2}(2\epsilon \mathrm{Re}\langle 0|E|1\rangle + \epsilon^2\langle 1|E|1\rangle)\,. \tag{5.82}$$

This value can be negative for small ϵ if $\mathrm{Re}\langle 0|E|1\rangle < 0$ which can easily be obtained by a particular choice of the phases.

Negative energy of this form is not sufficient to prevent singularities in general, but it does affect horizon properties. Black holes then become unstable and Hawking-radiate, which is just what one expects from the thermodynamical analogy. To make this quantitative, one has to compute the spectrum of the radiation and see whether it is thermal and at what temperature. This calculation requires quantum field theory on a curved spacetime. As originally used by Hawking (1971), on flat space one can define the usual notion of a vacuum state as the unique Poincaré-invariant state, based on the underlying symmetries of the background. If the background is different, however, there is no unique Fock representation of particle states. In fact, the concept of particles and antiparticles depends on the availability of a time translation symmetry to split waves into positive and negative frequency modes. This is not available on general curved spacetimes such as that of a black hole. Instead, one can only introduce asymptotic notions of incoming and outgoing particle states if spacetime is asymptotically flat.

One has the usual particle pictures on $\mathcal{J}^-$ and $\mathcal{J}^+$, but in general they do not match: what appears as the vacuum state on $\mathcal{J}^-$ looks, when evolved to $\mathcal{J}^+$, like an excited state. In a flat space, the field operator has a decomposition

$$\hat{\psi}(t,x) = \int \frac{\mathrm{d}^3k}{\sqrt{2\omega_k}}(e^{-i(\omega_k t - kx)}\hat{a}_k + e^{i(\omega_k t + kx)}\hat{a}_k^\dagger) \tag{5.83}$$

in field modes, where the annihilation operators $\hat{a}_k$ define the vacuum by the conditions $\hat{a}_k|0\rangle = 0$ for all k, and $\omega_k(k)$ is the dispersion relation of the field. If this is obtained in an asymptotically flat regime on a curved space-time, the asymptotic mode functions

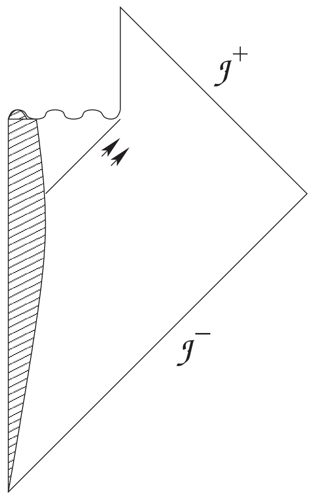

Fig. 5.29 A tentative conformal diagram describing the evoparation of a black hole by Hawking radiation. When the horizon disappears, a regular center may be left, but the part of space-time corresponding to the black-hole interior remains incomplete.

define the corresponding particle concept. But in different asymptotic regimes of a single space-time, such as $\mathcal{J}^-$ and $\mathcal{J}^+$, asymptotic expressions for the mode functions usually differ, mixing terms of $\hat{a}_k$ and $\hat{a}_k^\dagger$ when transforming from one regime to another one. In this formal sense, a past vacuum state can evolve into a future multi-particle state such as a thermal one.

If one uses an expectation value $\langle T_{ab}\rangle$ for a quantum field instead of a classical energy-momentum tensor, one can obtain a negative $R_{ab}\ell^a\ell^b$ along horizons. Then, the horizon becomes timelike and its area decreases. In this way, the thermodynamics of black holes becomes consistent when quantum effects are taken into account. Moreover, through the calculation of the temperature of Hawking radiation, all free constants are fixed unambiguously in terms of Planck's constant:

$$T_{\rm BH} = \frac{\hbar\kappa}{2\pi}\,, \qquad S_{\rm BH} = \frac{A}{4\hbar G}\,. \tag{5.84}$$

The combination $\hbar G$ determining the relation between area and entropy is the square of the Planck length.

Despite the consistency of this picture, black-hole evaporation is not completely understood. What is lacking is a common solution for radiating quantum fields together with their back-reaction on the underlying space-time geometry. Only such a solution, or at least a good approximation, can provide insights into the endstate of black-hole evaporation and show what conformal picture, possibly of the form shown in Fig. 5.29, is obtained. In this final stage of a black hole where the horizon falls into the strong curvature regime near the central singularity, moreover, it is no longer consistent to consider only quantized matter on a classical geometry. Here, a complete quantization including the gravitational field is required, giving rise to quantum gravity.

Exercises

5.1 (i) Derive an equation for the rate of change $\xi^c \nabla_c \omega_{ab}$ of the rotation of a timelike geodesic congruence with tangent vector field ξ^a.

(ii) Show that a timelike geodesic congruence which is initially hypersurface orthogonal remains so as long as its expansion and shear are finite.

5.2 Define a timelike geodesic congruence in Minkowski space by the straight lines generating the double cone $x^2 + y^2 + z^2 = \frac{1}{4}t^2$. Compute its expansion and verify that it diverges at the tip of the double cone.

5.3 Verify that also the expansion of the null geodesic congruence defined by generators of the double cone $x^2 + y^2 + z^2 = t^2$ in Minkowski space diverges at the tip.

5.4 Find the coordinate transformation $(t, r) \mapsto (T, r)$ that maps the Schwarzschild line element to Painlevé–Gullstrand form.

5.5 In a spherically symmetric region of space-time where the lapse function does not vanish, marginally trapped surfaces are given by solutions to $g^{rr} = 0$.

(i) Show that this condition can equivalently be written as

$$\frac{1}{g_{rr}} - \frac{(N^r)^2}{N^2} = 0\,. \tag{E5.1}$$

(ii) Evaluate (E5.1) for the Schwarzschild space-time in Painlevé–Gullstrand coordinates.

5.6 The Vaidya solution describing gravitational collapse is given by the line element

$$\mathrm{d}s^2 = -\left(1 - \frac{2M(v)}{r}\right)\mathrm{d}v^2 + 2\mathrm{d}v\mathrm{d}r + r^2\mathrm{d}\Omega^2$$

with a bounded, non-decreasing function $M(v) \geq 0$.

(i) Find the ingoing and outgoing null normal vector fields $k_\pm^a$ to a spherical surface defined by constant values of r and v. Check under which conditions they give tangent vectors of affinely parameterized geodesics, i.e. $k_\pm^a \nabla_a k_\pm^b = 0$. If they do not, rescale the normal vectors such that they give affinely parameterized future-pointing tangent vectors.

(ii) For a spherically symmetric null congruence, the normal tangent space up to multiples of $k_\pm^a$ is spanned by the angular vector fields $(\partial/\partial\vartheta)^a$ and $(\partial/\partial\varphi)^a$. Thus, $\hat{h}^{ab}$ is just the angular part of the inverse metric and the expansion of a spherical null congruence is given by

$$\theta = \hat{h}^{ab}\nabla_a k_b = g^{\vartheta\vartheta}\nabla_\vartheta k_\vartheta + g^{\varphi\varphi}\nabla_\varphi k_\varphi\,.$$

Compute this for the two normal congruences derived in part (i) and find the space-time region of the Vaidya solution which contains spherically symmetric trapped surfaces.

5.7 Use the Friedmann equation for positive spatial curvature to show that de Sitter space can be expressed as

$$ds^2 = -d\tau^2 + \alpha^2 \cosh^2(\tau/\alpha)(dr^2 + \sin^2 r\, d\Omega^2)$$

($0 \leq r < \pi$) and find the relation between α and the cosmological constant.

In order to find the conformal diagram of de Sitter space, introduce a new coordinate $t(\tau)$ (leaving r unchanged) such that the metric is conformal to the Einstein static universe. Derive and draw the resulting conformal diagram.

5.8 (i) Show that in complex coordinates $\zeta(\vartheta, \varphi) = e^{i\varphi} \cot \frac{\vartheta}{2}$ one has

$$d\zeta d\bar{\zeta} = \frac{1}{4}(1 + \zeta\bar{\zeta})^2 (d\vartheta^2 + \sin^2\vartheta\, d\varphi^2)\,.$$

(ii) For Cartesian coordinates x, y, z verify

$$x = r\frac{\zeta + \bar{\zeta}}{1 + \zeta\bar{\zeta}}\,, \qquad y = -ir\frac{\zeta - \bar{\zeta}}{1 + \zeta\bar{\zeta}}\,, \qquad z = r\frac{\zeta\bar{\zeta} - 1}{1 + \zeta\bar{\zeta}}$$

with $r^2 = x^2 + y^2 + z^2$ and polar angles ϑ, φ. Use this to show that ζ does not change up to order r^{-1} under a translation $x' = x + b$, $y' = y + c$, $z' = z + d$: $\zeta' = \zeta + O(r^{-1})$.

(iii) Consider now Minkowski space-time in Cartesian coordinates t, x, y, z and introduce the null coordinate $u = t - r$ where r is as before. Show that a translation $t' = t + a$ combined with the spatial translation of part (ii) corresponds to an asymptotic supertranslation of the form

$$u' = u + \frac{A + B\zeta + \bar{B}\bar{\zeta} + C\zeta\bar{\zeta}}{1 + \zeta\bar{\zeta}} + O(r^{-1})$$

with real A and C. Relate A, B and C to the translation parameters a, b, c, d.

5.9 Compute the ADM energy of a Schwarzschild black hole.

5.10 Compute the angular momentum of a Kerr black hole.

5.11 Show that the ADM mass equals the quasilocal expression for energy associated with a 2-sphere boundary of spatial slices, approaching the sphere at infinity.

5.12 (i) Perform the matching for an arbitrary isotropic interior region described by an FLRW model with scale factor $a(t)$ and an exterior region of generalized Vaidya form,

$$ds^2 = -\left(1 - \frac{2M(v, \chi)}{\chi}\right) dv^2 + 2dv d\chi + \chi^2 d\Omega^2\,.$$

Derive $M(v(t), \chi(t))$, $d\chi/dv$ and $\dot{v}$ along the matching surface in terms of the scale factor and its derivative, making sure that $\dot{v}$ is positive for a collapsing interior ($\dot{a} < 0$).

(ii) Matching vv-components of extrinsic curvature implies that

$$3(1-2\partial_\chi M)\frac{\mathrm{d}\chi}{\mathrm{d}v} = 2\left(\partial_v M + \chi\frac{\mathrm{d}^2\chi}{\mathrm{d}v^2}\right).$$

Use this to write $\partial_v M$ in terms of total derivatives of M and χ with respect to v along the matching surface $\chi(v)$.

(iii) Compute $\partial_v M$ at the place where the matching surface intersects the boundary $2M(v,\chi)=\chi$ of the region of spherical trapped surfaces. Total derivatives along the matching surface can be written in terms of time derivatives of the interior scale factor. Show that this boundary given by $2M(v,\chi)=\chi$ is always null where it intersects the matching surface.

5.13 Show that $\xi^a = v\left(\frac{\partial}{\partial v}\right)^a + \chi\left(\frac{\partial}{\partial\chi}\right)^a$ is a conformal Killing vector field of the Vaidya space-time

$$\mathrm{d}s^2 = -\left(1-\frac{2M(v)}{\chi}\right)\mathrm{d}v^2 + 2\mathrm{d}v\mathrm{d}\chi + \chi^2\mathrm{d}\Omega^2,$$

i.e. $\nabla_{(a}\xi_{b)} \propto g_{ab}$, if and only if the mass function is of the form $M(v)=\mu v$ with a constant μ.

6
Quantum gravity

In general relativity, the space-time metric provides the physical field of gravity and is subject to dynamical laws. For a complete and uniform fundamental description of nature, the gravitational force, and thus space-time, is to be quantized by implementing the usual features of quantum states, endowing it with quantum fluctuations and imposing the superposition principle. Only then do we obtain a fully consistent description of nature, since matter as well as the non-gravitational forces are quantum, described by quantum stress-energy which can couple to gravity only via some quantum version of the Einstein tensor.

An implementation of this program requires a clear distinction of the different concepts used in general relativity. One normally works with the line element for metric purposes, but this is a combination of metric tensor components and coordinate differentials (separating events from each other). Only the geometry is dynamical, not the coordinates. After quantization, we may have a representation for geometrical observables such as the sizes of physically characterized regions, but not for coordinates or distances between mathematical points as mere auxiliary ingredients. Metrics or other tensors may arise in an effective form from quantum gravity, but they are not the basic object. One has to dig deeper, similarly to hydrodynamics where the continuous fluid flow is not suitable for a fundamental quantum theory, which must, rather, be based on an atomic picture.

For quantum gravity, we do not know yet what fundamental structures are required, and we certainly do not have any observational hints. In such a situation, canonical methods come in handy, since they allow us to tell what the observable ingredients of a theory are in the sense of gauge invariance. For gravity, this means that we learn about the possibly observable signatures of space-time without having to refer to coordinates. We can express information contained in the geometry in a gauge-invariant way, at least in principle. Explicit constructions of gauge-invariant observables in general terms are certainly very difficult, constituting one of the major problems for quantum gravity. Nevertheless, many steps of constructing a canonical quantum theory of gravity can be followed, and they indicate key features for such formulations, as well as possible physical implications in cosmology and black-hole physics. In this chapter, only the basic outline of canonical quantum gravity will be given, with a focus on the canonical methods involved (rather than, for instance, issues of Hilbert space representations). The aim is to see how the gravitational dynamics is expected to change from quantum effects, and what implications it might have on global space-time issues. For more details of constructive canonical quantum gravity, we refer to the books by Kiefer (2004), Rovelli (2004) and Thiemann (2007). Different approaches to quantum gravity are described in the books by Oriti (2009) and Ellis et al. (2011).

6.1 Constrained quantization and background independence

As a constrained theory, general relativity requires several different constructions to be performed for a quantum formulation. For quantization in general terms, phase-space variables are to be turned into operators, which is a guideline with a clear aim even though no universal procedures exist to do that in practice. But in constrained theories, there are different kinds of phase space: the unrestricted (also called *kinematical*) one, on which constraint functions are defined, and the *reduced phase space* on which constraints have been solved and on which the flow of the first-class part has been factored out. If only a subset of the constraints is solved, there are other, in-between versions of phase spaces of different dimensions.

6.1.1 Reduced phase-space quantization

One can rightfully argue that the reduced phase space should be the most important one, since it contains only observable information without gauge freedom. All variables are restricted to the values allowed by the constraints, and by factoring out the gauge flow all configurations related by gauge have been identified as one single observable. One would solve the constraints before quantization, using just the classical calculations, and find a complete set of observables parameterizing the reduced phase space. The reduced phase space would then form the basis for constructing a quantum theory.

Example 6.1 (Quantized circle)
The constraint $C = (q^1)^2 + (q^2)^2 - R^2$ *on a 4-dimensional phase space* $\mathbb{R}^4$ *with canonical coordinates* (q^1, q^2, p_1, p_2) *constrains the two configuration variables to be on a circle of radius* R*. A local coordinate on the constraint surface is given by the angle* $\varphi = \arctan(q^2/q^1)$*, and it is easy to verify that a second independent observable is given by* $p_\varphi = q^1 p_2 - q^2 p_1$*, obeying* $\{p_\varphi, C\} = 0$*, which is conjugate to* φ*:* $\{\varphi, p_\varphi\} = 1$*. However, the expression for* φ *is only local and not suitable for a quantization; the reduced phase space is not* $\mathbb{R}^2$ *but the co-tangent space of the circle. A global parameterization cannot be given in terms of a single angle variable but rather, requires at least two coordinate charts.*

Instead of trying to implement coordinate transformations in quantum physics, a more successful strategy is to enlarge the number of observables considered. Not all of them will be independent, but they can all be globally defined. Specifically, we replace the single local function φ *by the globally defined* $\sin\varphi = q^2/\sqrt{(q^1)^2 + (q^2)^2}$ *and* $\cos\varphi = q^1/\sqrt{(q^1)^2 + (q^2)^2}$*. For the global variables, we obtain a new algebra under Poisson* *brackets:*

$$\{\sin\varphi, p_\varphi\} = \cos\varphi\,, \qquad \{\cos\varphi, p_\varphi\} = -\sin\varphi\,, \qquad \{\sin\varphi, \cos\varphi\} = 0 \tag{6.1}$$

which is not canonical but corresponds to the Lie algebra of the Euclidean group $\mathbb{R}^2 \rtimes$ SO(2) *of 2-space. Appropriate quantizations of the reduced phase space can be found from the representation theory of the Euclidean group (or its universal covering), using group theoretical methods as developed by Isham (1983) in general terms. No unique representation is picked out, and so a parameter due to quantization ambiguities arises even in the action of basic phase-space operators.*

The origin of the ambiguity parameter is easy to see in the angle representation of wave functions and basic operators. We choose states to be wave functions $\psi(\varphi)$*, providing a well-defined probability density if* $|\psi(\varphi)|^2$ *is periodic in* φ*. Representing the circle by the range* $0 \leq \varphi < 2\pi$*, the wave function must obey boundary conditions* $\psi(2\pi) = \psi(0)\exp(2\pi i\theta)$ *for some parameter* $\theta \in [0, 1)$*. Our Hilbert space is going to be the space of square-integrable functions on* $[0, 2\pi]$ *obeying the boundary conditions as specified with reference to* θ*. No boundary condition of this form is respected by* φ *as a multiplication operator, and so it cannot be represented. The global functions* $\sin\varphi$ *and* $\cos\varphi$*, on the other hand, have well-defined representations by multiplication. Together with* $\hat{p}_\varphi = -i\hbar \mathrm{d}/\mathrm{d}\varphi$*, the correct commutator representation of the classical Poisson relations (6.1) follows. For any given* θ*, we can choose an orthonormal basis of states* $\psi_n^{(\theta)}(\varphi) = \exp(i(n+\theta)\varphi)/\sqrt{2\pi}$*,* $n \in \mathbb{Z}$*. These states are eigenstates of the momentum operator* $\hat{p}_\varphi$ *with eigenvalues* $n + \theta$*. The discrete spectrum of the momentum on the circle is clearly affected by the ambiguity* θ*. Representations with different* θ *cannot be related by unitary transformations.*

This so-called reduced phase-space quantization is appealing due to its emphasis on gauge-invariant notions, but too often it turns out to be impractical. Solution spaces to the constraints may have complicated forms and topologies, difficult to be equipped with sufficiently simple phase-space variables to be turned into operators, with Poisson bracket relations (multiplied with $i\hbar$) becoming commutators. Even finite-dimensional systems of non-trivial topologies can be difficult to quantize; and if quantum representations can be found, they tend to be non-unique in those cases. Especially for general relativity, for which a complete set of observables to form the reduced phase space looks too difficult to be constructed explicitly, reduced phase-space quantization seems to be a limited option. (Still, model systems can sometimes be analyzed in this way, providing useful information.)

One may also worry that possibly crucial "off-shell" information would be overlooked: for quantum theories, configurations not solving the classical equations do contribute to the quantum theory, as can most intuitively be seen from a path-integral perspective. Constrained systems do not just have equations of motion but also their constraints, which follow from the action just as the evolution equations do. If one solves the constraints before quantization, one does not consider the full space of classical configurations and loses access to some of the off-shell part.

6.1.2 Dirac quantization

Several of the problems associated with reduced phase-space quantizations are avoided in the Dirac quantization procedure for constraints. Here, one turns the kinematical phase space, that on which the constraint functions are defined, into an operator algebra acting on the *kinematical Hilbert space*. With this algebra, one quantizes the constraints just as one would quantize the Hamiltonian of an unconstrained theory: one tries to find an operator with the classical expression as the correspondence. Once all the constraint operators $\hat{C}_I$ have been determined, they are solved by requiring *physical states* to be annihilated by all of them: $\hat{C}_I|\psi\rangle = 0$. Physical states thus solve the constraints, and they are automatically gauge-invariant, since the unitary flow $\exp(it\hat{C})$ generated by a self-adjoint constraint operator $\hat{C}$ acts as the identity on states annihilated by $\hat{C}$.

Example 6.2 (Circle)
Continuing with the preceding example, we easily represent the constraint as an operator in the standard position representation of the phase space $\mathbb{R}^4$. *The solution space requires physical states* $\psi(q^1, q^2) = \bar{\psi}(\varphi)\delta_{C=0}$ *to be supported only on the circle defined by the classical constraint, making solutions distributional and non-normalizable. The physical Hilbert space of states satisfying the Dirac constraints will be spanned by the functions* $\bar{\psi}$ *arising as coefficients of the delta function, but those states do not automatically come equipped with an inner product (as it were the case if solutions to constraints were normalizable, i.e. if zero were in the discrete spectrum of the constraint).*

Methods for inducing a physical inner product on the solution space exist, such as group averaging as described by Marolf (1995) and Giulini and Marolf (1999). One writes all physical states as $\bar{\psi} := \int \mathrm{d}t \exp(-it\hat{C})\psi$ *in terms of kinematical states* ψ. *In the present example, the integrated operator amounts to multiplication with a delta function on the circle. Instead of using the kinematical inner product for expectation values in* $\bar{\psi}$ *or overlaps with other states, which may not be well defined, a new physical inner product can be written as*

$$(\bar{\psi}_1, \bar{\psi}_2) = \bar{\psi}_1[\psi_2] := \int \mathrm{d}t \langle \psi_1, \exp(it\hat{C})\psi_2\rangle$$

using the evaluation of the distribution $\bar{\psi}_1$ *on the kinematical state* ψ_2 *that gives rise to* $\bar{\psi}_2$ *when group averaged. The evaluation is independent of the choice of* ψ_2 *as a kinematical representative of* $\bar{\psi}_2$ *because we are integrating over the full range of* t.

For the circle, we reproduce the sector $\theta = 0$ *in this way. The ambiguity is realized here by the fact that we could have chosen a different kinematical Hilbert space. If we had chosen to represent kinematical states not on the full plane but on the plane minus its origin, states periodic in the polar angle only up to a phase would be possible. Group averaging of those states leads to all the* θ*-sectors. These representations can be obtained only after removing the origin from the plane, and so one may prefer the* $\theta = 0$ *sector for which this removal is not required. However, the origin is an unphysical point of the* *constrained system, and so kinematical properties relying on its presence should not be used to distinguish representations.*

Also, this procedure is typically subject to quantization ambiguities, such as different choices to order the factors of a constraint operator. And although the kinematical phase space is often simple enough to be quantized by standard means, many mathematical subtleties can arise in determining physical states and the Hilbert space they form. If all physical states correspond to an eigenvalue of zero in the discrete part of the spectrum of all constraint operators, the physical Hilbert space is simply a subspace of the kinematical one, spanned by all the zero-eigenstates. But if this is not the case, not all physical states are kinematically normalizable; one will have to define a new, physical inner product on the solution space. Procedures for this exist, but that part is usually the most difficult one of a Dirac quantization in specific cases.

To extract information such as expectation values, one needs operators that are gauge-invariant in the sense that they commute with the constraint operators. Such operators $\hat{O}$ will map a physical state $|\psi\rangle$ to another one, $|\psi'\rangle = \hat{O}|\psi\rangle$: if $\hat{C}|\psi\rangle = 0$, $\hat{C}|\psi'\rangle = \hat{C}\hat{O}|\psi\rangle = \hat{O}\hat{C}|\psi\rangle = 0$ if $\hat{C}\hat{O} = \hat{O}\hat{C}$. At this stage, we are led back to the problem of observables. It is less severe than in the reduced phase-space quantization in some respects, since we do not require a complete set. If we are interested in a specific question, for instance regarding the evolution of a certain matter field in cosmology, one observable capturing this information would suffice. On the other hand, the problem of finding a single observable becomes more difficult than in a classical setting, since we must have two commuting operators $\hat{O}$ and $\hat{C}$ rather than a pair of phase-space functions with vanishing Poisson bracket. This is usually a more complicated problem because even the ordering of factors in the definition of the operator will matter.

A further crucial issue in Dirac quantizations is that of anomalies. As we have seen in several examples, the first-class nature of the Hamiltonian and diffeomorphism constraints of general relativity is not just important for the underlying gauge freedom but also for the consistency of all the equations involved. The same is true at the quantum level. Here, commutators $[\hat{C}_I, \hat{C}_J]$ of constraint operators vanish on states annihilated by the constraints, just as Poisson brackets $\{C_I, C_J\}$ must vanish on a first-class constraint surface.[1] The danger in quantizations, then, is that $[\hat{C}_I, \hat{C}_J]$ may not exactly equal the constraint introduced as the quantization $\{\widehat{C_I, C_J}\}$; there may be quantum corrections. In gravity, for instance, the commutator $[\hat{H}[N_1], \hat{H}[N_2]]$ would have to equal the quantized diffeomorphism constraint, including the quantized metric in its structure function. Only quantum corrections proportional to the constraints can be allowed in the commutator, so that they will vanish on-shell and do not destroy the first-class nature. If this is not the case, we would have two conditions, namely that $[\hat{C}_I, \hat{C}_J]$ as well as $\{\widehat{C_I, C_J}\}$ annihilate physical states, which can easily become inconsistent with each other and leave too small a solution space. In general

[1] Second-class constraints must be implemented differently: at the quantum level they cannot be solved sharply due to uncertainty relations. As non-commuting operators they would not have a common eigenspace of zero-eigenstates. Normally one solves second-class constraints classically before quantization, but other methods exist as discussed by Klauder (2001).

situations, specifying suitable classes of semiclassical states sufficiently explicitly can be a further challenge.

Example 6.3 (Anomalies)
The constraints $C_1 := q^1$ and $C_2 := q^2 + q^1 p_1^2$ form a first-class set. When we quantize the system, factor-ordering choices arise in $\hat{C}_2$, for which we may choose $\hat{C}_2^{(1)} = \hat{q}^2 + \hat{q}^1 \hat{p}_1^2$, $\hat{C}_2^{(2)} = \hat{q}^2 + \hat{p}_1^2 \hat{q}^1$ or (perhaps preferably) the symmetric $\hat{C}_2^{(3)} = \hat{q}^2 + \hat{p}_1 \hat{q}^1 \hat{p}_1$. Combinations of these versions are also possible, such as another symmetric option $\frac{1}{2}(\hat{C}_2^{(1)} + \hat{C}_2^{(2)})$.

For these choices, the algebra of constraint operators will be

$$[\hat{C}_1, \hat{C}_2^{(1)}] = 2\hat{q}^1 \hat{p}_1 \,, \qquad [\hat{C}_1, \hat{C}_2^{(2)}] = 2\hat{p}_1 \hat{q}^1 \,, \qquad [\hat{C}_1, \hat{C}_2^{(3)}] = \hat{p}_1 \hat{q}^1 + \hat{q}^1 \hat{p}_1$$

using $\hat{C}_1 = \hat{q}^1$. Only $\hat{C}_2^{(2)}$ provides an anomaly-free quantization, since $[\hat{C}_1, \hat{C}_2^{(2)}]$ annihilates all states annihilated by $\hat{C}_1$. Moreover, the anomalous terms $\hat{q}^1 \hat{p}_1$ in the other commutators are not even defined on states annihilated by $\hat{C}_1$. We can order $\hat{q}^1$ to the right, $\hat{q}^1 \hat{p}_1 = \hat{p}_1 \hat{q}^1 + i\hbar$, in order to have it act on states first, and annihilate physical ones. But this will give rise to non-zero (and imaginary) reordering terms. Due to the anomaly, the quantum system in those cases is overconstrained and leaves no physical solutions at all.

In particular, no symmetric ordering of $\hat{C}_2$ produces an anomaly-free system. Using $\hat{C}_1^{(2)}$, there is an anomaly-free quantization with a non-symmetric constraint operator. Expectation values of the constraint are then not guaranteed to be real, even though the classical constraint is real. But this is acceptable given that the constraint is not an observable, and that the reality statement at this level refers to the kinematical inner product, which is anyway going to change on the physical Hilbert space. Making sure that basic observables become self-adjoint operators can be used as a guiding principle to determine a physical inner product. For gravity, the necessity of non-Hermitian constraints has been argued for by Komar (1979).

If an anomaly-free quantization can be found and interesting quantum observables are available, the last issue is the selection of a suitable class of states to be studied. For a comparison with the classical behavior, these states should be semiclassical: they should provide expectation values near to what we would expect from the classical observables, but still show deviations as they may arise from quantum fluctuations.

6.1.3 Effective space-times

At least the problem of observables can be evaded by using effective constraints. For an anomaly-free algebra of constraint operators and a class of kinematical semiclassical states (not annihilated by the constraints), effective constraints are defined as expectation values of the constraint operators in these states, as well as of products of other operators with a constraint as the rightmost factor. For information about observables, the effective constraints then still have to be solved, which requires extra constructions as we will provide them below. But even before doing so, the effective constraint functionals themselves, and the algebra they obey under Poisson brackets, provide interesting insights. Effective constraints also make many of the classical canonical techniques of Chapter 3 available for quantum systems.

Example 6.4 (Expectation values of constraints)
In the preceding example, all physical states must have vanishing expectation values $\langle\hat{q}^1\rangle = 0$, $\langle\hat{q}^2 + \hat{p}_1^2\hat{q}^1\rangle = 0$ *with the constraints. Moreover, the expectation value* $\langle\hat{p}_1^2\hat{q}^1\rangle$ *contained in the expectation value of* $\hat{C}_2^{(2)}$ *automatically vanishes if* $\hat{C}_1$ *is already imposed. Thus, in the anomaly-free quantization restrictions on expectation values of the basic operators* $\langle\hat{q}^1\rangle$ *and* $\langle\hat{q}^2\rangle$ *are consistent.*

With the anomalous quantization $\hat{C}_2^{(3)}$*, by contrast, the expectation value*

$$0 = \langle\hat{C}_2^{(3)}\rangle = \langle\hat{q}^2\rangle + \langle\hat{p}_1\hat{q}^1\hat{p}_1\rangle = \langle\hat{q}^2\rangle + \langle\hat{p}_1^2\hat{q}^1\rangle + i\hbar\langle\hat{p}_1\rangle\,,$$

if it can consistently vanish in an implementation of the quantum constraint, requires $\langle\hat{q}^2\rangle = -i\hbar\langle\hat{p}_1\rangle$*. This relation, following from the anomaly, is problematic for two reasons: we expect* $\langle\hat{q}^2\rangle$ *to be a physical observable (its classical analog has vanishing Poisson brackets with all constraints), and so it should be real. This can be realized only for imaginary values of* $\langle\hat{p}^1\rangle$*. Moreover, the physical variable* $\langle\hat{q}^2\rangle$ *refers to the non-physical* $\langle\hat{p}^1\rangle$ *which is subject to gauge transformations generated by* $\hat{C}_1$*.*

For an anomaly-free quantization, the effective constraints are first class. But their phase-space functionals as well as the algebra they form typically differ from the classical constraints by quantum corrections, providing what is called a consistent deformation of the classical theory: the equations of the classical theory are modified — “deformed” — but in a consistent way without breaking the number of gauge transformations. The form of gauge transformations and the observables they imply may change, but the transformation algebra after deformation is as large as the classical one. No gauge transformation is broken or violated, and no first-class constraint becomes second class.

In some cases, deformed constraints may still satisfy the classical algebra. As computed in detail by Deruelle et al. (2009), this happens, for instance, if quantum effects imply the addition of higher-curvature terms to the Einstein–Hilbert action. While the dynamics changes in this case, the hypersurface-deformation algebra remains of the classical form. This robustness of the algebra is not surprising, for the algebra contains very basic information about the deformation of hypersurfaces in a space-time manifold, which is still expected to be realized for higher-curvature theories.

But examples also exist in which even the constraint algebra is deformed, and structure functions are quantum corrected. In such a case, transformations generated by the constraints can no longer be interpreted as space-time diffeomorphisms. We can still transform the full space-time metric using the quantum-corrected effective constraints and their algebra, as discussed in general terms in Chapter 3.3.4.3 and used explicitly for classical cosmological perturbations in Chapter 4.4.2.2. While observables can still be computed and interpreted consistently in this way, metric information cannot be expressed in the form of an "effective" line element: if gauge transformations of metric components are no longer of the form of space-time diffeomorphisms as indicated by a deformed algebra, they differ from the transformations of coordinate differentials. Line elements constructed in the usual way, multiplying dynamical solutions for metric components with coordinate differentials, would not be gauge invariant under the corrected transformations. Such line elements, making use of the ordinary notion of coordinates, do not correspond to observable information, and results would depend on the space-time gauge chosen. In fact, if the underlying gauge transformations are not space-time diffeomorphisms, we cannot use the usual space-time manifold structure to interpret our observables geometrically. Generalized versions are, for instance, provided by non-commutative spaces and their geometries, on which the deformed diffeomorphisms would act.

Explicit instances of such consequences can usually be found in simplified settings such as symmetry reduced models or phenomenologically motivated effective equations. We will show examples in the following sections on quantum cosmology and black holes, but first finish the discussion of general constructions with further specifics about the basic operators of candidate quantum theories of gravity.

6.1.4 Kinematical operators

Following the Dirac procedure, even before constraints are quantized and imposed, the kinematical setting of quantum gravity provides subtleties. Kinematical quantization means that the basic canonical phase-space variables, such as h_{ab} and p^{ab} in the ADM formulation of general relativity, are to be turned into operators such that their classical Poisson algebra is represented by the commutator algebra. In other words, the procedure of quantum mechanics, turning position q and momentum p into operators $\hat{q}$ and $\hat{p}$ such that $[\hat{q}, \hat{p}] = i\hbar\widehat{\{q, p\}} = i\hbar$, is to be extended to general relativity.

6.1.4.1 Scalar field

This extension is challenging. First, the classical Poisson algebra of field theories is not strictly an algebra: we have $\{h_{ab}(x), p^{cd}(y)\} = \delta^c_{(a}\delta^d_{b)}\delta(x, y)$ as a distribution, not a function. The normal procedure to arrive at a well-defined algebra free of infinite coefficients is *smearing*. For a scalar field, for instance, the ill-defined algebra $\{\varphi(x), p_\varphi(y)\} = \delta(x, y)$ can be made well defined by using the integrated fields $\varphi[N] := \int d^3x N(x)\sqrt{\det h}\varphi(x)$ and $p_\varphi[M] := \int d^3x M(x) p_\varphi(x)$ with $\{\varphi[N], p_\varphi[M]\} = \int d^3x \sqrt{\det h} N(x)M(x)$. (With p_φ being a scalar density, no metric factor is required to make its integration well defined.)

Here, a new problem arises: integrating the fields with scalar multiplier functions requires a background metric at least for some of them, φ in the case considered so far. Otherwise, the integration would not be well defined in all cases. Making use of a background metric is acceptable if a field theory is to be quantized on a given space-time manifold, such as electromagnetism on a possibly curved space-time. A metric would readily be available, and not be quantized itself.

If we quantize gravity, however, the metric is one of the objects to be turned into an operator. As a field, it must be smeared, requiring an integration measure. We could use an auxiliary metric to define the measure for smearing the actual physical metric, but this background structure would remain in the theory and is likely to lead to spurious results. Or, we could use the physical metric to define the integration measure for its own smearing. But then, the smeared metric would be highly non-linear and not even polynomial, preventing the existence of a simple Poisson algebra of basic objects to be turned into an operator algebra.

To resolve this situation, one must look for alternative smearings. For the scalar field this is easy. Instead of using an ordinary scalar smearing function $N(x)$, we could use a scalar density. With this understanding, $\varphi[N] = \int d^3x N(x)\varphi(x)$ is well defined without any metric factor, and so is $\{\varphi[N], p_\varphi[M]\} = \int d^3x N(x)M(x)$. Or, we could simply drop the smearing of φ while using the previous one for p_φ, resulting in the well-defined $\{\varphi(x), p_\varphi[M]\} = M(x)$.

6.1.4.2 Gravity

A similar route can also be taken for gravity, although the choices are more involved due to the tensorial nature of fields. Historically, the first quantization, due to Wheeler and DeWitt, turned the classical fields into formal operators acting on wave functionals $\psi[h_{ab}]$ by multiplication with $\hat{h}_{ab}$ and functional derivatives of $\hat{p}^{ab} = -i\hbar\delta/\delta h_{ab}$, without worrying about the smearing. Accordingly, this quantization has remained purely formal except for homogeneous models of quantum cosmology in which only finitely many degrees of freedom remain, as introduced by DeWitt (1967). Another difficulty is that the space of all metrics, on which wave functions are defined, is hard to control or to equip with a good measure to be used for normalizing wave functions.

Loop quantum gravity has taken the smearing issue more seriously, and found a solution, first proposed by Rovelli and Smolin (1990) for gravity (but preceded by Gambini and Trias (1980) for non-gravitational interactions). It turns out that connection variables are much better to control than metric variables, which helps in finding a measure to normalize states defined as wave functions of connections rather than metrics. Moreover, a natural background-independent smearing is available. Now we are dealing with fields given by a connection A_a^i (the Ashtekar–Barbero[2] connection) and a densitized triad E_j^b, such that $\{A_a^i(x), E_j^b(y)\} = 8\pi\gamma G\delta_a^b\delta_j^i\delta(x, y)$. The connection can easily be integrated 1-dimensionally along curves e in space, by defining holonomies

$$h_e(A) = \mathcal{P}\exp\left(\int_e ds^a A_a^i\tau_i\right). \tag{6.2}$$

Here, $\tau_j = -\frac{1}{2}i\sigma_j$, with Pauli matrices σ_j, are generators of the Lie algebra su(2), such that holonomies take values in SU(2). The symbol $\mathcal{P}$ indicates that the ordering of factors

[2] Initially, complex connections were used which simplify the constraints. But these variables come with two major difficulties: complicated reality conditions would have to be imposed after quantization, and the non-compact structure group obtained by complexifying SU(2) makes defining inner products of states challenging.

in the matrix exponential, defined by its power series expansion, must be specified suitably along the path. Holonomies then satisfy the parallel-transport equation

$$\frac{\mathrm{d}h_e(\lambda)}{\mathrm{d}\lambda} = A_a^i \dot{e}^a \tau_i h_e(\lambda) \tag{6.3}$$

for the family $h_e(\lambda)$ of holonomies along pieces of the path e between parameter values 0 and λ.

Similarly, a densitized triad can be naturally integrated 2-dimensionally over surfaces S, resulting in fluxes

$$F_S^{(f)}(E) = \int_S \mathrm{d}^2y E_i^a n_a f^i \tag{6.4}$$

with an su(2)-valued smearing function f^i. Here, the co-normal n_a to the surface does not require an auxiliary metric, unlike the normal. (If the surface is specified as a constant-level surface $g(x) = \text{const}$, for instance, $n_a = \partial_a g$.)

Holonomies and fluxes are suitable phase-space variables, since A_a^i and E_j^b can be recovered from them when computed for all curves and surfaces, as shown explicitly for holonomies by Giles (1981). Curves and surfaces (and matrix indices) have thus replaced the labels a, i and x of the initial fields. Most importantly, the Poisson algebra between holonomies and fluxes is well defined, as explicitly computed by Lewandowski et al. (1993): the integrations combine to a 3-dimensional one, integrating out the delta-function of the field algebra and resulting in the intersection number of a curve and a surface used.

To quantize these objects, one may work in a connection representation and view holonomies as the analog of creation and annihilation operators. Starting from a simple state $\psi[A_a^i] = 1$, we construct more complicated ones by multiplying with matrix elements of holonomies along collections of curves e_I:

$$\psi_{\{e_I, A_I, B_I\}}[A_a^i] = \prod_I \left(h_{e_I}(A_a^i)\right)^{A_I}_{B_I} . \tag{6.5}$$

Although for any finite collection of edges states depend on the connection only along 1-dimensional graphs $g = \{e_I\}$ defined by the curves used, it turns out that the set of all states is large enough to represent all connections. Moreover, taking all finite graphs, we obtain an orthonormal basis of a Hilbert space which endows holonomies with the right adjointness properties. As defined by Ashtekar and Lewandowski (1995), the inner product can be written as

$$\langle \psi, \phi \rangle = \int_{\mathrm{SU}(2)^n} \prod_{I=1}^{n} \mathrm{d}\mu_{\mathrm{Haar}}(h_I) \psi(h_1, \ldots, h_n)^* \phi(h_1, \ldots, h_n) \tag{6.6}$$

for two states based on the same graph with edges e_I, $I = 1, \ldots, n$ (and a vanishing inner product for states based on different graphs). Here, $\mathrm{d}\mu_{\mathrm{Haar}}$ is the normalized Haar measure on SU(2), endowing the group with unit volume. At this place the compactness of the group matters: the inner product is required for states on an infinite-dimensional field space, and so for general states we have to integrate over all possible curves in space. For states depending only on a finite number of holonomies, almost all the integrations reduce to unit factors. If the structure group were non-compact, for instance as a consequence of complexification, it would not support a normalizable invariant measure and reductions to finite dimensions would be impossible.

Flux operators in the connection representation act as functional derivative operators

$$\hat{F}_S^{(f)} = \frac{8\pi\gamma G\hbar}{i} \int_S \mathrm{d}^2 y n_a f^i \frac{\delta}{\delta A_a^i(y)} .$$

The product $G\hbar = \ell_{\mathrm{P}}^2$ is the square of the Planck length. Acting on a state of the form (6.5), we have

$$\hat{F}_S^{(f)} \psi_{\{e_I, A_I, B_I\}} = -8i\pi\gamma\ell_{\mathrm{P}}^2 \sum_I \int_S \mathrm{d}^2 y \int_{e_I} \mathrm{d}\lambda n_a(y) \dot{e}^a \delta(y, e(\lambda)) f^i (\tau_i)_{B_I}^{A_I} \cdot \psi_{\{e_I, A_I, B_I\}} \quad (6.7)$$

showing that the graph states are eigenstates of fluxes. The eigenvalues depend on the intersection numbers $\mathrm{Int}(S, e) = \int_S \mathrm{d}^2 y \int_{e_I} \mathrm{d}\lambda n_a(y) \dot{e}^a \delta(y, e(\lambda))$ between the surface through which we compute the flux and the curves in the graph, which takes integer values. Also, the remaining coefficients are discrete, and so flux operators have discrete spectra.

Since fluxes quantize the densitized triad, related to the spatial metric, a discrete version of quantum geometry is indicated. In fact, from the fluxes one can, following Rovelli and Smolin (1995) and Ashtekar and Lewandowski (1997, 1998), construct area operators by quantizing $A_S = \int_S \mathrm{d}^2 y \sqrt{n_a E_i^a n_b E_i^b}$ or volume operators by quantizing $V_R = \int_R \mathrm{d}^3 x \sqrt{|\det(E_j^b)|}$. And also, these operators have discrete spectra, so that spatial geometry is indeed realized in a quantized way. This does not yet tell us what space-time looks like on small scales, for we still have to implement the constraints. But spatial discreteness is one important geometrical property that can be expected to have dynamical implications. It is also one of the key features by which loop quantum gravity deviates from a Wheeler–DeWitt quantization even in its basic representation. All this is a rather direct consequence of background independence which has forced us to consider objects such as holonomies and fluxes, and to construct their representations.

6.2 Quantum cosmology

If one considers isotropic space-times or the homogeneous ones of Bianchi type, just a finite number of phase-space parameters suffices to characterize the metric. In such a reduced setting, there is a finite number of degrees of freedom; we are no longer dealing with a field theory and no smearing is required to represent the basic variables. One might expect that the key differences between Wheeler–DeWitt quantum gravity and loop quantum gravity, caused mainly by a different viewpoint toward the smearings involved, would disappear. This, however, turns out not to be the case, and so quantum cosmology provides an interesting setting to explore and test these frameworks.

6.2.1 Representation

A Wheeler–DeWitt quantization of isotropic cosmological models initially looks just like quantum mechanics. As seen in Chapter 2, we have a single pair of gravitational phase-space variables, a and p_a, plus possible matter variables. For quantization, in this context one usually chooses the metric representation, with wave functions $\psi(a)$ depending on the scale factor, normalized by $\int \mathrm{d}a|\psi(a)|^2 = 1$. Accordingly, $\hat{a}$ is a multiplication operator and $\hat{p}_a = -i\hbar\partial/\partial a$. Clearly, $\hat{a}$ has a continuous spectrum and there is no sign of discrete spatial geometry.

6.2.1.1 Physical Hilbert spaces

A subtlety concerns the self-adjointness of the basic operators, especially the momentum. If $a > 0$ as required classically, we are quantizing a system with a coordinate restricted to the positive half-line. While $\hat{a}$ is self-adjoint, $\hat{p}_a$ is not. This can be seen by several means: first, $\hat{p}_a$ generates a translation in a by the operator $\exp(i\epsilon\hat{p}_a)$; shifting $\psi(a)$ out of its domain of definition when $a = 0$ is crossed means that $\exp(i\epsilon\hat{p}_a)$ cannot be unitary. Second, $\hat{p}_a$ has an eigenstate $\psi(a) = c\exp(-a)$, normalizable on the half-line, belonging to a non-real eigenvalue. One can deal with this issue by using more suitable operator algebras as developed under the name affine quantum gravity by Klauder (2003, 2006). Alternatively, one may take the viewpoint that only self-adjointness relations of physical operators, those quantizing observables defined on the physical phase space solving the constraints, are relevant.

In any case, the next step is to raise the classical constraints to conditions for states in the Hilbert space of quantum cosmology. As in quantum mechanics, this is possible by replacing any a by $\hat{a}$ and any p_a by $\hat{p}_a$ in the classical constraint (2.15) (as well, possibly, as matter components by their quantizations). Thus, one obtains a constraint operator

$$\hat{C} = -\frac{2\pi G}{3}\frac{\hat{p}_a^2}{\hat{a}} - \frac{3}{8\pi G}k\hat{a} + \frac{\hat{H}_{\mathrm{matter}}}{N}\,. \tag{6.8}$$

Due to products of the non-commuting $\hat{a}$ and $\hat{p}_a$, there is no unique definition for $\hat{C}$.

Physical states are those annihilated by the constraint: $\hat{C}|\psi\rangle = 0$. If zero is contained in the discrete spectrum of $\hat{C}$, there are solutions in the original, kinematical Hilbert space whose inner product we can use for the solution space. For zero in the continuous part of the spectrum, on the other hand, physical states are not normalizable by the kinematical inner product; a new physical Hilbert space must be determined in this case. As mentioned in the general discussion of the preceding section, several procedures exist, such as group averaging, guided by the requirement that observables must become self-adjoint operators on the physical Hilbert space. For most of these procedures, $\hat{C}$ must be self-adjoint.

The condition for physical states solves the constraint. Its classical analog is first class, which means that a complete solution requires not just finding the constraint surface but also factoring out the gauge. For the quantization of a first-class constraint, no gauge need be factored out. In fact, the transformation generated by the constraint via the unitary operator

$\exp(i\epsilon\hat{C})$ (replacing the Poisson flow generated by a classical constraint) is already reduced to the identity for physical states annihilated by $\hat{C}$. The condition of physical states is thus the only one implementing a first-class constraint, making the imposition of first-class quantum constraints rather natural.

Example 6.5 (Physical Hilbert space of deparameterizable systems)
In this example we consider systems with two degrees of freedom t and q with momenta p_t and p, subject to one deparameterizable constraint. Before implementing the constraint, we use the kinematical representation of wave functions $\psi(t, q)$.

For a non-relativistically deparameterizable constraint with quantization $\hat{C} = \hat{p}_t + \hat{H}(q, p)$, the constraint equation $\hat{C}|\psi\rangle = 0$ requires that $\psi(t, q)$ satisfy the Schrödinger equation with Hamiltonian $\hat{H}$. Physical states are uniquely characterized by their "initial" values $\psi(t_0, q)$ at some fixed t_0. An inner product on physical states is easily defined as $\langle\psi_1, \psi_2\rangle = \int \mathrm{d}q\,\psi_1^*(t_0, q)\psi_2(t_0, q)$, making $\hat{q}$ and $\hat{p}$ self-adjoint operators. The physical inner product is independent of the choice of t_0 for a self-adjoint Hamiltonian $\hat{H}$, generating unitary evolution.

For a relativistic deparameterization, the constraint equation $(\hat{p}_t^2 - \hat{H}^2)|\psi\rangle = 0$ implies the Klein–Gordon-type equation $(-\hbar^2\partial_t^2\psi - \hat{H}^2)\psi = 0$ for physical states. Solutions to this equation can be obtained from solutions to the Schrödinger equation $i\hbar\partial_t\psi = \pm|\hat{H}|\psi$, so-called positive and negative frequency solutions, or superpositions thereof. A conserved bilinear form now is of the form

$$(\psi_1, \psi_2) = i\int \mathrm{d}q(\psi_1^*\partial_t\psi_2 - \psi_2\partial_t\psi_1^*) \tag{6.9}$$

but it is not positive definite: it produces positive norms of positive-frequency solutions and negative norms of negative-frequency solutions. Since positive and negative-frequency solutions of the same absolute value of the frequency are automatically orthogonal, a slight modification of the conserved bilinear form to $\langle\psi_1, \psi_2\rangle := (\psi_1, \psi_2)$ if both states are positive frequency, $\langle\psi_1, \psi_2\rangle := -(\psi_1, \psi_2)$ if they are both negative frequency, and $\langle\psi_1, \psi_2\rangle := 0$ in the mixed case produces a valid inner product. The relationship of this inner product with group averaging has been explored by Hartle and Marolf (1997).

If a complete eigenbasis of $|\hat{H}|$ with eigenstates φ_n and eigenvalues $\hbar\omega_n$ is known, positive-frequency and negative-frequency physical states can be expanded as

$$\psi_\pm(t, q) = \sum_n c_n\varphi_n(q)\exp(\mp i\omega_n t)$$

if $|\hat{H}|$ has a discrete spectrum (and replacing the sum by an integration otherwise). For a free, massless relativistic particle, for instance, $\hat{H} = \hat{p}$ with eigenstates $\varphi_k(q) = \exp(ikq)$, k real. And $|\hat{H}|$, by definition of the operator absolute value, has the same eigenstates φ_k with eigenvalues $\hbar|k|$.

6.2.1.2 Loop representation

Loop quantum cosmology, reviewed by Bojowald (2008, 2011), implements the same steps as loop quantum gravity, but in a reduced setting. Its results can rather easily be compared with those of a Wheeler–DeWitt quantization, showing key differences between the strategies. For isotropic models, only holonomies evaluated in isotropic connections $A_a^i = \tilde{c}\delta_a^i$ appear; see Example 4.13. Along straight lines in the direction of translation symmetries $X_I^a = (\partial/\partial x^I)^a$, holonomies $\exp(\int X_I^a A_a^i \tau_i)$ in the fundamental representation of SU(2) have matrix elements of the form $\exp(i\mu c)$ where μ depends on the length of the curve used. Here, it turns out to be useful to introduce $c := V_0^{1/3}\tilde{c}$ in terms of the coordinate size V_0 of the region used to define the isotropic phase space as in Example 4.13, and so μ for a given curve also depends on V_0. As in the general situation of loop quantum gravity, we start with a simple state in the connection representation, $\psi(c) = 1$, and create more complicated ones by multiplying with holonomies. This setup of the construction is thus very different from Wheeler–DeWitt quantum cosmology. While Wheeler–DeWitt quantization uses quantum mechanical intuition to define basic operators and the kinematical Hilbert space, loop quantum cosmology uses tools developed for full quantum gravity based on the principle of background independence. Several ingredients of the kinematical construction can be induced from those in the full theory. In the full theory, uniqueness theorems for the holonomy-flux representation have been proved by Lewandowski et al. (2006) and Fleischhack (2009), which by the induction distinguishes the analogous properties of the models.

All states obtained in this way can be written as

$$\psi_{\{\mu_I, f_I\}}(c) = \sum_I f_I \exp(i\mu_I c)\,,$$

with an inner product

$$\langle\psi_1, \psi_2\rangle = \lim_{T\to\infty} \int_{-T}^{T} \mathrm{d}c \psi_1^* \psi_2\,. \tag{6.10}$$

For general graph states (6.5) without homogeneity, we obtained orthogonal ones when different graphs were used. Here similarly, we obtain orthogonal states when different $\{\mu_I\}$ are used. Thus, the set $\{\exp(i\mu_I c) : \mu_I \in \mathbb{R}\}$ is a basis of our resulting Hilbert space. Moreover, unlike plane waves in quantum mechanics, the states $\psi_{\mu_I}(c) := \exp(i\mu_I c)$ are normalized.

What is constructed in this way is a representation inequivalent to the one used by Wheeler and DeWitt. There is no unitary transformation connecting the two representations, seen by the fact that they differ in properties which would be preserved by any unitary transformation. For instance, holonomy operators quantizing $\exp(i\mu c)$ by the construction sketched above are not continuous in μ: they have matrix elements

$$(\psi_{\mu_1}, \exp(i\mu c)\psi_{\mu_2}) = \delta_{\mu_1,\mu+\mu_2} = \begin{cases} 1 & \text{if } \mu = \mu_1 - \mu_2 \\ 0 & \text{otherwise.} \end{cases}$$

Thus, while the exponentials are well-defined operators, no operator for c can be derived by taking a derivative of $\widehat{\exp(i\mu c)}$ by μ. In a Wheeler–DeWitt quantization, by contrast, an operator $\hat{c}$ could easily (though not uniquely, owing to factor ordering choices) be obtained from $\hat{a}$ and $\hat{p}_a$ using $\tilde{c} = \gamma\dot{a} \propto p_a/a$ for the example of spatially flat models.

In order to extract geometrical information, we use triad operators. An isotropic densitized triad has the form $E^a_i = \tilde{p}\delta^a_i$, where $|\tilde{p}| = a^2$ can be read off from the relation to the spatial metric. Analogously to c, we define $p := V_0^{2/3}\tilde{p}$. (Since the triad can have two different orientations, p can take either sign. This information is not contained in the metric. Based on triad variables, loop quantum cosmology thus easily avoids the question of how to quantize a system on the positive half-line, were we to use a metric representation. With the metrical variable p allowed to take all real values, we quantize on the full real line.) As seen in Example 4.13, it turns out that p is canonically conjugate to c, $\{c, p\} = 8\pi\gamma G/3$, and so p becomes a derivative operator $\hat{p} = -\frac{8}{3}i\pi\gamma\ell_{\rm P}^2\partial/\partial c$. On the basis states $\psi_\mu(c) = \exp(i\mu c)$, it acts by

$$\hat{p}\psi_\mu = \frac{8\pi\gamma}{3}\ell_{\rm P}^2\mu\psi_\mu$$

clearly showing the triad spectrum. (Now we can see why the redefinitions from $(\tilde{c}, \tilde{p})$ to (c, p) were useful: the Poisson bracket and quantum representations do not explicitly depend on V_0. The specific powers of V_0 are motivated by the dimensionality of smearings to holonomies and fluxes.)

Eigenvalues of $\hat{p}$ can take any real numbers, suggesting a continuous spectrum. However, the eigenstates are normalizable, which is normally a characteristic of states in the discrete spectrum of an operator. The apparent contradiction is resolved if we observe that the Hilbert space with which we are dealing here is not separable: it has an uncountable basis $\{\exp(i\mu c) : \mu \in \mathbb{R}\}$. In this situation, normalizability of eigenstates turns out to be the more crucial requirement for a discrete spectrum, and so fluxes are discrete in loop quantum cosmology. (The real line can be equipped with continuous or discrete topologies, and so the set of eigenvalues forming the real line does not tell us much about discreteness unless we know the relevant topology. Using the normalizability of states as a criterion is free of this limitation.) Similarly, the spatial volume $\hat{V} = |\hat{p}|^{3/2}$ of the integration region of coordinate size V_0 has a discrete spectrum.

Discreteness and continuity properties, or the lack thereof, clearly show the differences in representations. Also, dynamics notices these features. In Wheeler–DeWitt quantum cosmology one can directly quantize the Hamiltonian constraint, resulting in a second-order differential operator. In loop quantum cosmology, by contrast, there is no operator directly for c; only exponentials $\exp(i\delta c)$ are represented. Since these are non-linear expressions in c, higher powers of c, or of curvature components, must necessarily be included in such a quantization. Additional choices are now required, such as the value (or even p-dependence; see Chapter 6.2.3) of δ used and the precise form of a function agreeing with c^2 at small curvature. For instance,

$$\hat{C}_{\text{loop}} = -\frac{3}{8\pi G}\frac{\widehat{\sin(\delta c)}^2}{\gamma^2\delta^2}\sqrt{|\hat{p}|} + \hat{H}_{\text{matter}} \tag{6.11}$$

has a well-defined representation while the original classical constraint containing a term with c^2 has not. This kind of regularization is not unique, but in any case the resulting constraint equation for a state ψ_μ in the triad representation is a difference rather than differential operator, since $\exp(i\delta c)$ acts on states by a finite shift of the triad eigenvalue. At small volume and large curvature, the discreteness matters and leads to different behaviors of wave functions.

6.2.2 Effective treatment and implications

Implications of mathematical differences in the dynamics are best shown by effective equations, which implement the key quantum effects but do not require us to deal directly with wave functions. Such formulations can be done rather similarly to classical canonical ones, except that quantum degrees of freedom appear together with the classical ones. The material briefly reviewed in this section presents another application of Hamiltonian concepts.

6.2.2.1 Quantum phase space

The commutator of operators on a Hilbert space satisfies almost all the relations required for a Poisson bracket — antisymmetry, linearity and the Jacobi identity — but it is not defined on a function space and thus lacks an analog of the Leibniz rule. A complete link between commutator algebras and Poisson structures can easily be established by defining

$$\{\langle\hat{A}\rangle, \langle\hat{B}\rangle\} := \frac{\langle[\hat{A}, \hat{B}]\rangle}{i\hbar} \tag{6.12}$$

where $\langle\hat{A}\rangle$, for an arbitrary fixed operator, is interpreted as a (densely defined) function on the space of states used to compute the expectation value: $\langle\hat{A}\rangle(\psi) = \langle\psi|\hat{A}|\psi\rangle$ for a pure state or more generally $\langle\hat{A}\rangle(\rho) = \mathrm{tr}(\hat{A}\rho)$. By linearity and the Leibniz rule, the definition (6.12) can easily be extended to arbitrary combinations of the expectation values. The state space thus becomes a phase space with a Poisson bracket for functions defined on it.

Given a canonical pair $\hat{a}$ and $\hat{p}_a$ of basic operators, the newly defined Poisson bracket mimicks the classical bracket: $\{\langle\hat{a}\rangle, \langle\hat{p}_a\rangle\} = 1$. Nevertheless, the quantum Poisson algebra is a major extension of the classical one, for it applies to infinitely many more degrees of freedom. Not only expectation values amount to independent information about a state, but with $\langle\hat{a}^2\rangle \neq \langle\hat{a}\rangle^2$ in general, quantum fluctuations such as $(\Delta a)^2 = \langle(\hat{a} - \langle\hat{a}\rangle)^2\rangle$ provide independent variables. In fact, there are infinitely many new variables, since any product of arbitrarily many factors of $\hat{a} - \langle\hat{a}\rangle$ and $\hat{p}_a - \langle\hat{p}_a\rangle$ results in independent moments

$$G^{\overbrace{a\cdots a}^{k}\overbrace{p_a\cdots p_a}^{l}} = \langle(\hat{a} - \langle\hat{a}\rangle)^k(\hat{p}_a - \langle\hat{p}_a\rangle)^l\rangle_{\text{Weyl}} \tag{6.13}$$

with totally symmetric ordering (indicated by the subscript "Weyl"). The quantum state space is infinite-dimensional, and so clearly requires infinitely many independent parameters.

Moments are linear combinations of integer powers of expectation values taken for products of operators, functions to which our Poisson bracket can be applied. It then follows that

$$\{\langle\hat{a}\rangle, G^{\overbrace{a\cdots a}^{k}\overbrace{p_a\cdots p_a}^{l}}\} = 0 = \{\langle\hat{p}_a\rangle, G^{\overbrace{a\cdots a}^{k}\overbrace{p_a\cdots p_a}^{l}}\}$$

for all moments, and so the quantum variables supply a subspace symplectically orthogonal to the classical one.

Example 6.6 (Second-order moments)
At second order, we have three independent moments: two fluctuations $G^{aa} = \langle(\hat{a} - \langle\hat{a}\rangle)^2\rangle = (\Delta a)^2$ *and* $G^{p_a p_a} = \langle(\hat{p}_a - \langle\hat{p}_a\rangle)^2\rangle = (\Delta p_a)^2$ *as well as the covariance* $G^{ap_a} = \frac{1}{2}\langle\hat{a}\hat{p}_a + \hat{p}_a\hat{a}\rangle - \langle\hat{a}\rangle\langle\hat{p}_a\rangle$. *Poisson brackets are*

$$\{G^{aa}, G^{p_a p_a}\} = \{\langle\hat{a}^2\rangle - \langle\hat{a}\rangle^2, G^{p_a p_a}\} = \frac{1}{i\hbar}\langle[\hat{a}^2, \hat{p}_a^2]\rangle - 2\{\langle\hat{a}^2\rangle, \langle\hat{p}_a\rangle\} = 4G^{ap_a}$$
$$\{G^{aa}, G^{ap_a}\} = 2G^{aa}$$
$$\{G^{ap_a}, G^{p_a p_a}\} = 2G^{p_a p_a}\,.$$

On the quantum phase space of expectation values and all moments, the expectation value of the Hamiltonian operator defines a Hamiltonian flow equivalent to what one would obtain from solutions of the Schrödinger equation. For most systems, this flow couples all infinitely many moments to each other and to the expectation values. This is, in fact, how quantum corrections arise: moments such as fluctuations couple to expectation values, and so the changing shape of a wave function influences its mean position. Exact solutions of such infinitely coupled systems are difficult to find, but there are approximation schemes, for instance an adiabaticity assumption exploited by Bojowald and Skirzewski (2006) for anharmonic systems, as well as rare, exactly solvable models in which the equations decouple. We will first illustrate the application to constrained systems, developed by Bojowald et al. (2009a) and Bojowald and Tsobanjan (2009), and then discuss a solvable model for quantum cosmology.

Example 6.7 (Effective constraints for the free, massless particle)
The constraint operator $\hat{C} = \hat{p}_t^2 - \hat{p}^2$ *implies the effective constraint*

$$C := \langle\hat{C}\rangle = \langle\hat{p}_t\rangle^2 - \langle\hat{p}\rangle^2 + (\Delta p_t)^2 - (\Delta p)^2 = 0\,,$$

which must vanish when evaluated in a physical state, but also requires expectation values such as

$$C_{p_t} := \langle(\hat{p}_t - \langle\hat{p}_t\rangle)\hat{C}\rangle = 2\langle\hat{p}_t\rangle(\Delta p_t)^2 - 2\langle\hat{p}\rangle G^{p_t p} = 0$$
$$C_p := \langle(\hat{p} - \langle\hat{p}\rangle)\hat{C}\rangle = 2\langle\hat{p}_t\rangle G^{p_t p} - 2\langle\hat{p}\rangle(\Delta p)^2 = 0$$

(computed up to moments of third order) to vanish since $\hat{C}$ is acting on the state to the right. In a semiclassical state, moments of third (or higher) order are suppressed compared to those of second order by an additional power of $\hbar$, and so solving constraints for second-order moments and expectation values can provide a good approximation to the full quantum behavior. A finite-dimensional constrained system results that can be analyzed by standard means; there is, however, one difference compared to usual canonical systems in classical physics because the phase-space of second-order moments is 3-dimensional: it cannot be a symplectic manifold but is outright Poisson. This feature sometimes requires care in the discussion of gauge flows and the classification of first-class versus second-class constrained surfaces (see the Appendix).

In the present example, we solve $C_p = 0$ to obtain $G^{p_t p} = (\Delta p)^2\langle\hat{p}\rangle/\langle\hat{p}_t\rangle$, and with this solve $C_{p_t} = 0$ for $(\Delta p_t)^2 = G^{p_t p}\langle\hat{p}\rangle/\langle\hat{p}_t\rangle = (\Delta p)^2\langle\hat{p}\rangle^2/\langle\hat{p}_t\rangle^2$. The first constraint $C = 0$ then leads to a fourth-order equation

$$\langle\hat{p}_t\rangle^4 - (\langle\hat{p}\rangle^2 - (\Delta p)^2)\langle\hat{p}_t\rangle^2 - \langle\hat{p}\rangle^2(\Delta p)^2 = 0$$

for the expectation value of $\hat{p}_t$, solved by $\langle\hat{p}_t\rangle = \pm\langle\hat{p}\rangle$ just as classically. Had we included a non-trivial potential, additional moments would have appeared in the solution, implying deviations from the classical behavior.

For observables, the gauge flow of effective constraints on the quantum phase space is to be considered. If the system is deparameterizable with t as time, the structure of the constraints shows that all moments involving $\hat{t}$ can be fixed as part of the gauge choice, while moments involving $\hat{p}_t$ are determined by solving the constraints. Only moments not involving $\hat{t}$ or $\hat{p}_t$ remain dynamical and free to choose as initial values. Solving the constraints for the expectation value of $\hat{p}_t$, an expression of the form $\langle\hat{p}_t\rangle = H_{\rm eff}(\langle\hat{q}\rangle, \langle\hat{p}\rangle, (\Delta q)^2, (\Delta p)^2, G^{qp}, \ldots)$ results correcting the classical deparameterized Hamiltonian by moment terms. This is the same form as expected for a quantization performed after deparameterization.

6.2.2.2 Solvable models

A free, massless scalar field in isotropic cosmology can serve well for deparameterization, since the only term its energy density contributes is kinetic and does not contain φ. For a spatially flat isotropic model in connection variables, we have the Hamiltonian constraint

$$-\frac{3}{8\pi G}\frac{c^2}{\gamma^2}\sqrt{|p|} + \frac{1}{2}\frac{p_\varphi^2}{|p|^{3/2}} = 0$$

which can be reformulated in deparameterized form as

$$p_\varphi \pm \sqrt{\frac{3}{4\pi G}}\frac{|cp|}{\gamma} = 0\,.$$

For a Hamiltonian of the form $\hat{H} = \pm\frac{1}{2}\sqrt{3/4\pi\gamma^2 G}\,(\hat{c}\hat{p} + \hat{p}\hat{c})$ we can easily compute the expectation value $H_Q = \langle\hat{H}\rangle = \pm\sqrt{3/4\pi\gamma^2 G}(\langle\hat{c}\rangle\langle\hat{p}\rangle + G^{cp})$. (We have dropped the norm, which seems dangerous but is justified for states supported only on the positive part of the spectrum of $\hat{c}\hat{p} + \hat{p}\hat{c}$.) The quantum Hamiltonian generates the flow $\mathrm{d}\langle\hat{c}\rangle/\mathrm{d}\varphi = \pm 2\sqrt{4\pi G/3}\,\langle\hat{c}\rangle$ and $\mathrm{d}\langle\hat{p}\rangle/\mathrm{d}\varphi = \mp 2\sqrt{4\pi G/3}\,\langle\hat{p}\rangle$ (independent of any moments for this solvable model) which is easily solved by exponentials. The densitized-triad expectation value $\langle\hat{p}\rangle$ reaches zero asymptotically, where the energy density and curvature diverge. Although it takes an infinite amount of φ to evolve there, it is a finite amount of proper time. To see this, we just need to relate proper time τ to φ by solving $\mathrm{d}\varphi/\mathrm{d}\tau = |p|^{-3/2} p_\varphi$, using the fact that p_φ is a constant of motion. The model thus remains singular after a Wheeler–DeWitt quantization.

In loop quantum cosmology, the Hamiltonian must change, since holonomy operators, exponentials of c, are represented but not c itself. This automatically introduces higher-order terms in c, and the system is no longer solvable as before. A Hamiltonian constraint that can be represented is, for instance,

$$-\frac{3}{8\pi G}\frac{\widehat{\sin^2(\delta c)}}{\gamma^2\delta^2}\sqrt{|\hat{p}|} + \frac{1}{2}\widehat{\frac{1}{|p|^{3/2}}} p_\varphi^2 = 0 \tag{6.14}$$

obtained from (6.11) with the Hamiltonian of a free, massless scalar. This expression shows the higher-order terms when Taylor expanded in c, as a consequence of holonomy corrections. Also present is a second type of correction from the discrete nature of quantum geometry: flux operators, just as $\hat{p}$ in the isotropic reduction, have discrete spectra containing the eigenvalue zero, and so they cannot have densely defined inverse operators. On the other hand, an operator whose classical limit is $|p|^{-3/2}$ is needed to quantize the matter Hamiltonian and analyze the quantum constraint. Such operators exist despite the non-existence of a precise inverse operator: if we write

$$\frac{1}{|p|^{3/2}} = \left|\frac{3}{4\pi\gamma G}\{c, \sqrt{|p|}\}\right|^3 = \left|\frac{3i}{4\pi\gamma\delta G}e^{i\delta c}\{e^{-i\delta c}, \sqrt{|p|}\}\right|^3$$

following a general procedure of loop quantum gravity developed by Thiemann (1998b,c) and then quantize the right-hand side using holonomy operators, the flux operator for the positive power $\sqrt{|p|}$, and then turning the Poisson bracket into a commutator by $i\hbar$, we obtain a well-defined operator

$$\widehat{\frac{1}{|p|^{3/2}}} = \left|\frac{3}{8\pi\gamma\delta\ell_{\mathrm{P}}^2}\left(\widehat{e^{i\delta c}}\left[\widehat{e^{-i\delta c}}\sqrt{|\hat{p}|}\right] - \widehat{e^{-i\delta c}}\left[\widehat{e^{i\delta c}}, \sqrt{|\hat{p}|}\right]\right)\right|^3 .$$

Its eigenvalues

$$\left(\frac{3}{8\pi\gamma\ell_{\mathrm{P}}^2}\right)\delta^{-3}\left|\sqrt{\mu+\delta} - \sqrt{\mu-\delta}\right|^3 \sim \left(\frac{8\pi\gamma\ell_{\mathrm{P}}^2|\mu|}{3}\right)^3 ,$$

obtained in the triad eigenstates, agree with the expected result for $|p|^{-3/2}$ for $\mu \gg \delta$ but show strong deviations for small μ. In particular, they are regular for all μ, automatically cutting off the classical divergence at $p = 0$.

Both corrections — from holonomies and from the inverse triad — have been analyzed in several models. Holonomy corrections are more interesting in the present context because they allow us to keep the model of solvable form even in the presence of quantum effects. Even with the higher-order terms implied by non-polynomial expressions in c, it turns out that, in a certain ordering of the factors, the quantized model displays the same kind of solvability in new variables. (Moreover, as first seen qualitatively by Date and Hossain (2005) and then numerically by Ashtekar et al. (2006), this type of correction in the given model provides an interesting example of bouncing solutions in cosmology. The relevance and different types of bouncing cosmology have been reviewed by Novello and Bergliaffa (2008).)

If, following Bojowald (2007), we introduce the variables (p, J) with $J := p\exp(i\delta c)$ instead of the canonical ones (c, p), we can realize the φ-Hamiltonian corresponding to (6.14) in linear form,

$$p_\varphi = H = \pm\sqrt{\frac{3}{4\pi G}}\frac{|\mathrm{Im}J|}{\gamma\delta}\,. \tag{6.15}$$

By itself a linear Hamiltonian in some variables does not guarantee solvability, but here it does because the basic variables (p, J) form a linear algebra:

$$\{p, J\} = \frac{8\pi\gamma G}{3}i\delta J\,, \qquad \{p, \bar{J}\} = -\frac{8\pi\gamma G}{3}i\delta\bar{J}\,, \qquad \{J, \bar{J}\} = -\frac{16\pi\gamma G}{3}i\delta p\,.$$

We include $\bar{J}$ as an independent variable in addition to the complex J. The variables will eventually have to be cut back to two real ones by imposing the reality condition $J\bar{J} = p^2$, or its quantized analog.

6.2.2.3 Cosmological bounces

For a linear Hamiltonian (6.15), the effective Hamiltonian $H_Q = \langle\hat{H}\rangle$ equals the classical one. Thus, quantum equations for expectation values of $\hat{p}$ and $\hat{J}$ are identical to the classical ones,

$$\frac{\mathrm{d}\langle\hat{p}\rangle}{\mathrm{d}\varphi} = 2\sqrt{\frac{4\pi G}{3}}\mathrm{Re}\langle\hat{J}\rangle\,, \qquad \frac{\mathrm{d}\langle\hat{J}\rangle}{\mathrm{d}\varphi} = 2\sqrt{\frac{4\pi G}{3}}\,\langle\hat{p}\rangle\,.$$

Solving these equations gives

$$\langle\hat{p}\rangle(\varphi) = A\exp(2\sqrt{4\pi G/3}\varphi) + B\exp(-2\sqrt{4\pi G/3}\varphi)$$

$$(\langle\hat{J}\rangle\varphi) = A\exp(2\sqrt{4\pi G/3}\varphi) - B\exp(-2\sqrt{4\pi G/3}\varphi) + i\sqrt{\frac{4\pi G}{3}}\gamma\delta p_\varphi$$

where the imaginary part of $\langle\hat{J}\rangle$ is determined by the φ-Hamiltonian in terms of p_φ.

We see that $\langle\hat{p}\rangle(\varphi)$, compared to the Wheeler–DeWitt solutions, has a different mixture of exponentials. At this stage there might still be a singularity of $\langle\hat{p}\rangle(\varphi)$ being zero, which will happen at some φ when the integration constants A and B have opposite signs, but we have not yet imposed the reality condition. At the quantum level, requiring $\hat{J}\hat{J}^\dagger = \hat{p}^2$, this is done by imposing $\langle \hat{J} \rangle \overline{\langle \hat{J} \rangle} - \langle \hat{p} \rangle^2 = O(\hbar)$ which need not be sharply zero, since fluctuation terms would add to the expectation value of the quadratic condition. But as one can verify by analyzing the evolution of moments, the extra terms are constants of motion. They are thus small for a state required to be semiclassical at just one value of φ which could be at large volume — a reasonable assumption. Then, we have

$$-4AB + \frac{4\pi G}{3}\gamma^2\delta^2 p_\varphi^2 = O(\hbar)$$

and $AB > 0$. With A and B required to have the same sign by the reality condition, $\langle \hat{p} \rangle(\varphi)$ never reaches zero, not even asymptotically, and the singularity is resolved. In this model, the classical singularity is replaced by a bounce: a non-zero minimum of the volume. Although the high symmetry and the specific matter content make this model very special, it illustrates how the general results of singularity theorems can be avoided by quantum effects. Even though we did not change the classical matter contribution, thus not violating energy conditions, the gravitational dynamics changes. In particular, the Raychaudhuri equation no longer applies in the classical form, which was used to relate caustics of geodesics to incompleteness. How this can be realized at the level of sufficiently general space-times remains an open problem.

Other systems will be more complicated, since quantum back-reaction effects — the coupling of fluctuations, correlations or higher moments of a state to its expectation values — does arise and can in general not be ignored. The structure and fate of the singularity might then be more complicated to discuss at least in a long-term picture, but the approach to a quantum singularity can be analyzed in terms of the higher-dimensional dynamical systems provided by expectation values with a certain number of moments. At this stage, the analysis proceeds in a way close to canonical gravity, just with more variables and possibly extra contributions such as reality conditions of partially complex variables or uncertainty relations for some of the moments.

6.2.3 Lattice refinement

Another issue to be taken into account for more realistic modelling is lattice refinement. We have seen that loop quantum cosmology replaces the continuous geometry of classical gravity and of the Wheeler–DeWitt quantization by a discrete structure. States in a symmetry-reduced model are supposed to model behavior of a full state when averaged over large distances. Classically, only the scale factor and its momentum (plus matter variables) remain for an isotropic model. But in a discrete version of quantum gravity one would expect that the local change of the discreteness size, which is not fixed by the Planck length but can vary according to the excitation level of states, provides an independent evolving variable. It is not easy to implement such changing lattices in the context of quantum field theory, as discussed in several different contexts, e.g. by Weiss (1985), Unruh (1997) and Jacobson (2000).

Indeed, several models of loop quantum cosmology have shown that dynamical refinement must happen for consistency with semiclassical behavior at large volume: when a universe expands, its underlying discrete structure of space must be refined so as to avoid expanding the elementary scales to macroscopic sizes. Ensuring this by a dynamical mechanism is one of the key consistency requirements on discrete versions of quantum gravity. In full generality, no such mechanism has been formulated. But in homogeneous constructions one can at least model different refinement behaviors and see if and how consistent pictures can arise. Stability considerations have been discussed in this context by Bojowald et al. (2007), and phenomenological constraints found by Nelson and Sakellariadou (2007a,b) and Grain et al. (2010).

At the basic level of loop quantum cosmology, lattice refinement affects the spacing of the dynamical difference equation. Difference operators arose from the action of isotropic holonomies $\exp(i\delta c)$ so that δ determines the step size of the equation. Geometrically, δ encodes the length of a (straight) curve used to compute the holonomy for an isotropic connection. If lattice refinement happens, curves in a graph underlying a state will have to change when parameterized in a fixed coordinate system; to embed more curves in a region of constant coordinate size V_0, coordinate lengths of all curves must shrink. While the classical expansion is described by an increasing scale factor $a(\varphi)$ with respect to internal time φ, lattice refinement shrinks the curves and thus requires an a-dependent parameter $\delta(a)$.

When used in a Hamiltonian constraint operator, the step size of the difference equation becomes non-uniform: it depends on the step. In isotropic models, one can change variables so as to make the equation uniform again, at least up to factor-ordering terms. But this will no longer be possible in anisotropic models where more complicated versions of the equations occur.

Lattice refinement also has an influence on the exactly solvable model where $\exp(i\delta(p)c)$ must be used in the new variable J if we still want to produce a linear Hamiltonian. But this will change the algebra between J and $\bar{J}$, making it in general non-linear (and not even periodic in c). For power-law behaviors $\delta(p) = \delta_0 p^x$ with some parameter x, however, we can perform a canonical transformation $p' := p^{1-x}/(1-x)$, $c' := p^x c$ and use (p', J') with $J' := p' \exp(i\delta_0 c')$ in the same way as before. None of the follow-up equations will change, and the dynamics of the solvable model is insensitive to the refinement scheme as long as it is power-law. Non-solvable models, however, depend more sensitively on the refinement.

6.2.4 Quantum-cosmological perturbations

Homogeneous models often show possible implications of quantum effects in direct ways, but they may be too special for general conclusions. Reliable statements can be made only if results have been shown to have the potential of remaining intact when at least linear perturbations by inhomogeneities are included. For quantum cosmology, this also provides access to cosmological structure formation.

Once inhomogeneity is included, the constraint algebra becomes a major point to discuss. In a homogeneous model, there is just one constraint, the Hamiltonian constraint for constant lapse, which by necessity has a vanishing Poisson bracket with itself. At this level, quantum corrections can be included nearly at will, with possible restrictions only from consistency considerations of the quantum representations used. The main examples we have seen are the modification of $\dot{a}^2 \propto c^2$ in the isotropic constraint by periodic functions representable on the Hilbert space of loop quantum cosmology, and the behavior of inverse-triad expressions.

For inhomogeneous constraints, even if they are formulated for linearized equations, the situation changes drastically. The anomaly problem immediately rises to full strength, requiring all quantum corrections to be such that constraints remain first class even in their modified form. If this is not realized, evolution equations will not be of purely gauge-invariant form, coupling gauge artefacts to gauge-invariant perturbations. Corrections in perturbation equations must be tightly related to corrections in background equations for gauge-dependent terms to decouple from the physical dynamics. The equations may also be inconsistent if anomalies are present; for instance, from the gravitational equations for metric perturbations one would derive a second-order evolution equation for the matter source in conflict with the matter evolution equation. In covariant terms, this would correspond to mismatched corrections in the Einstein and stress-energy tensors such that they no longer fulfill the same form of conservation equations.

One may avoid discussing the constraint algebra by fixing the gauge before quantum corrections are computed, or by adopting a reduced phase-space approach where only the gauge-invariant variables would be quantized, as advocated for cosmology, e.g., by Giesel et al. (2007). But none of these procedures appears generic enough to be applicable in all necessary situations, and to show all quantum effects.

In loop quantum cosmology, consistent cosmological perturbation equations arising from an anomaly-free system of constraints have been found. But there is as yet no general formulation including all the expected quantum corrections. In particular with holonomy corrections, crucial to produce bouncing solutions in solvable models, consistent deformations turn out to be challenging to find explicitly.

Another correction, resulting from changes in the behavior of inverses of densitized-triad components, has a simpler form and has been included in consistent deformations found by Bojowald et al. (2008). Its analysis shows key implications of possible quantum-gravitational effects in cosmological structure formation, and also for the fundamental structure of quantum space-time. In particular, the constraint algebra is indeed deformed: it remains first class but has quantum corrections in its structure functions.

This result may be surprising, for quantum gravity corrections are often expected to be of higher-curvature type, but all higher-curvature actions give rise to the same form of the constraint algebra. The algebra, after all, describes the deformation of hypersurfaces in space-time, as they are used for any canonical formulation. If the hypersurface-deformation algebra changes, one can only conclude that the elementary structure of space-time, even as a manifold, is replaced by a different object in quantum gravity. And even semiclassical gravity will have a remnant of this new structure in the corrected algebra of its effective constraints. What this new manifold structure is, perhaps of non-commutative form, remains an open question. Changes to space-time structures are also of potential significance for cosmological observations, for they may affect the general classical results about conservation of power on large scales seen in Chapter 4.4.4.1. If power is not exactly preserved, even if only by a small amount, corrections may add up during long evolution times from Hubble exit to re-entrance. In this way, the potential for quantum-gravity corrections is larger than expected: higher-curvature corrections in cosmology would give rise to extra terms about the size of the tiny $\ell_P\mathcal{H}$, but additional magnification effects can occur during long-term evolution if the form of conservation laws changes.

6.3 Quantum black holes

Black holes provide another class of examples for solutions of general relativity in which singularities occur; they thus present additional tests of quantum gravity. But by necessity they are less symmetric than isotropic space-times, and thus more challenging.

As seen in Chapter 5.2.1, the region inside the horizon of a Schwarzschild black hole is homogeneous because the Killing vector field which is timelike outside the horizon turns spacelike and combines with the rotational symmetries to a transitive group action. The interior resembles a homogeneous cosmological model and its singularity may be eliminated by the same mechanisms, as discussed by Ashtekar and Bojowald (2006) and Modesto (2006). Results are so far less specific, since a homogeneous model can be used only for vacuum black holes, lacking any matter that could provide high densities at which a bounce of the geometrical variables could occur. A possible way out may be provided by matching techniques applied to non-singular matter models as collapsing isotropic star interiors. But in order to analyze black-hole singularities in general, including the vacuum case, one has to enter much more into a discussion of strong quantum regimes.

Qualitatively at least, one can speculate what a non-singular quantum extension of a classical black-hole space-time might look like. Especially the horizon behavior can change dramatically because quantum effects (even those of matter alone) easily violate energy conditions; there is no reason to assume horizons to be strictly spacelike or null in strong quantum regimes. Horizons may shrink from Hawking radiation, or possibly disappear completely owing to stronger quantum-gravity effects. If this were to happen, the black-hole interior would be opened up, no longer covered by a horizon. Light would have been trapped only for a possibly long but finite duration, but not forever. Trapping horizons would still form, and with them the usual black-hole properties. But if they disappear later on, the total space-time would be complete; no event horizon would exist. With horizons allowed to become timelike by quantum effects, smooth effective geometries can become possible in which an inner and outer horizon in a low-curvature regime close up at strong curvature to form a joint horizon enclosing the black-hole region from all sides. Based on negative-energy results, such behaviors have been proposed by Frolov and Vilkovisky (1981), Roman and Bergmann (1983) and Hayward (2006).

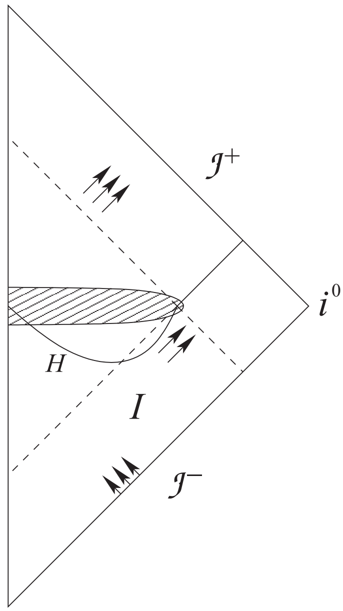

Fig. 6.1 A non-singular diagram of black-hole evaporation. A trapping horizon H forms as it would classically, but then dissolves in the strong quantum-gravity regime (hashed).

With quantum-gravity effects describing the collapse of an isotropic matter configuration, as realized for isotropic models of quantum cosmology, matching techniques can show what implications this might have for an effective space-time surrounding a star. Constructions follow the classical case of Oppenheimer–Snyder, except that the modified interior dynamics can no longer be matched to a static exterior; quantum-gravity corrections lead to effective pressures. Instead, a more general matching to Vaidya-type space-times can be constructed, showing the geometry at least in a neighborhood of the matching surface. (Only in a neighborhood because the matching to a general class of solutions does not determine functions such as $M(\chi, v)$ everywhere in a generalized-Vaidya space-time.) Following Bojowald et al. (2005a), one concludes that the initial collapse proceeds as it does classically, with a trapping horizon forming and enclosing the collapsing matter at some time. Differences to the classical case arise at high densitities, where the cosmological model used for the interior may show a bounce. The radius of the matter region as a function of time is no longer monotonic, and a second intersection of the surface with a trapping horizon can arise. Looking at the detailed equations, one finds that the matching breaks down beyond that point; physically, the collapsed interior would reappear from within the horizon, not describable by a simple matching. No complete constructions of non-singular collapse models exist yet, but this picture of horizons is consistent with the expectation of a closed one, extending the two intersections at the matching surface as well as the inner horizon formed within the matter region. As discussed by Ashtekar and Bojowald (2005) and illustrated in Fig. 6.1, several features of the endpoint of black-hole evaporation change in this context.

Further properties of black holes whose potential corrections in canonical quantum gravity are being studied are critical phenomena in the collapse process, found classically by Choptuik (1993) and analyzed with quantum-geometry effects by Husain (2009), and deviations from the thermality of Hawking radiation, e.g. by Vachaspati and Stojkovic (2008). Finally, by providing further specific examples for consistent deformations of the classical constrained system of general relativity, black-hole space-times with spherical symmetry, such as the constraints analyzed by Bojowald et al. (2009c), are providing means to probe the quantum structure of space-time.

6.4 The status of canonical quantum gravity

Here, we have only provided a sketch of the basic principles, methods and some applications of quantum gravity related to the canonical formulation. As indicated, promising effects have been found which can solve conceptual problems or may show the way to observational tests, but many open problems remain. Within this setting, canonical methods are indispensable owing to the tight control they give over observables, gauge properties and space-time structure.

Appendix A

Some mathematical methods

AA.1 Lie derivatives

Given a vector field v on a manifold M, one can take derivatives of tensor fields on M along its direction. Unlike the covariant derivative, the definition does not require any extra structures such as a connection or a metric. For a scalar α, the Lie derivative is equivalent to the definition of a vector field as a derivation: $\mathcal{L}_v\alpha = v\alpha = v^a\partial_a\alpha$. Once the Lie derivative of a vector field along another vector field is defined, the definition can be extended to all tensor fields using the Leibniz rule.

It turns out that the vector field commutator $\mathcal{L}_v w = [v, w]$ with $[v, w]^a = v^b\partial_b w^a - w^b\partial_b v^a$ satisfies the requirements for a derivation acting on w and can thus be used as a suitable definition for the Lie derivative: $\mathcal{L}_v(fw^a) = f\mathcal{L}_v w^a + (vf)w^a$ for a function f. For 1-forms ω_a, for instance, we use the Leibniz rule and the scalar nature of $\omega_a w^a$ with an arbitrary vector field w^a. Thus,

$$\mathcal{L}_v(\omega_a w^a) = (\mathcal{L}_v\omega_a)w^a + \omega_a[v, w]^a = (\mathcal{L}_v\omega_a)w^a + \omega_a v^b\partial_b w^a - \omega_a w^b\partial_b v^a$$

on the one hand and

$$\mathcal{L}_v(\omega_a w^a) = v^a\partial_a(\omega_b w^b) = v^a(\partial_a\omega_b)w^b + \omega_b v^a\partial_a w^b$$

on the other implies

$$\mathcal{L}_v\omega_a = v^b\partial_b\omega_a + \omega_b\partial_a v^b$$

since w^a was arbitrary.

The concept of the Lie derivative does in fact follow from a general viewpoint, applying to all tensor fields. We start with the vector field v^a along which the derivative is to be defined, and consider the 1-parameter family of diffeomorphisms $\Phi_t^{(v)}\colon M \to M$ it defines by integration. (This is, in fact, a 1-parameter group satisfying $\Phi_t^{(v)} \circ \Phi_s^{(v)} = \Phi_{t+s}^{(v)}$.) Thus, $\Phi_0^{(v)}$ is the identity and $\mathrm{d}\Phi_t^{(v)}/\mathrm{d}t|_p$ at each point p is identical to the vector field $v(p)$ at this point. Here, the t-derivative is interpreted as the vector field acting on functions α by

$$\left.\frac{\mathrm{d}\Phi_t^{(v)}}{\mathrm{d}t}\right|_p \alpha = \left.\frac{\mathrm{d}\alpha(\Phi_t^{(v)}(p))}{\mathrm{d}t}\right|_{t=0} . \tag{A.1}$$

The 1-parameter family of diffeomorphisms can be found by integrating ordinary first-order differential equations in coordinates.

The 1-parameter family of diffeomorphisms Φ_t, by its action on M, defines a pull-back action on functions and covariant tensors as well as a push-forward on contravariant tensors. The pull-back on functions is simply defined by $\Phi_t^*\alpha(p) = \alpha(\Phi_t(p))$. On vector fields w, the push-forward $\Phi_{t*}w$ is defined by mapping the integral curves of the vector field with the diffeomorphism and then computing the vector field along the new integral curves. A covariant tensor field ω_a is pulled back by $(\Phi_t^*\omega_a)w^a = \omega_a(\Phi_{t*}w^a)$ for any w^a. Extensions to higher tensor fields then follow from usual tensorization. (The distinction between push-forward and pull-back becomes clearer if one considers the more general case of maps from one manifold M to another one, N. Then, the push-forward of a contravariant tensor on M is a tensor of the same type on N, while the pull-back of a covariant tensor on N is a tensor of the same type on M. Unless the mapping from M to N considered is invertible, it does not give rise to a mapping of covariant tensors from M to N, or of contravariant ones from N to M.)

By push-forward with Φ_t or pull-back with $\Phi_t^{-1} = \Phi_{-t}$, the 1-parameter family of diffeomorphisms can be used to define the general Lie derivative along the vector field v whose integration gives Φ_t. On an arbitrary tensor T of some type, Φ_t defines a t-dependent mapping which we simply denote by $\Phi_t T$. Then, we have

$$\mathcal{L}_v T = \frac{\mathrm{d}\Phi_t^{(v)} T}{\mathrm{d}t} . \tag{A.2}$$

As one can verify by direct calculation, this provides the formulas for Lie derivatives already given for scalars and vector fields. For scalars, this is (A.1). For vector fields, we work in a small neighborhood around some point p and use local expressions $v = \sum_i v^i \partial/\partial x^i$ and $w = \sum_i w^i \partial/\partial x^i$. Applied to a coordinate function x^i in the neighborhood,

$$(\Phi_{t*}^{(v)} w(\Phi_t^{(v)}(p)))x^i = w(\Phi_t^{(v)}(p))(x_i \circ \Phi_{-t}^{(v)}) = \sum_j w^j(\Phi_t^{(v)}(p)) \left(\delta_j^i - t\frac{\partial g_{(v)}^i}{\partial x^j}\right)$$

where $g_{(v)}^i(t,q)$ is defined such that the flow for small t is described by $x^i(\Phi_{-t}^{(v)}(q)) - x^i(q) = -tg_{(v)}^i(t,q)$. (Thus, $g_{(v)}^i(0,q) = v^i(q)$.) For the Lie derivative, we need

$$\frac{1}{t}\left((\Phi_{t*}^{(v)} w)x^i - wx^i\right) = \frac{w^i(\Phi_t^{(v)}(p)) - w^i(p)}{t} - \sum_j w^j(\Phi_t^{(v)}(p))\frac{\partial g_{(v)}^i(t,\Phi_t^{(v)}(p))}{\partial x^j}$$

which for $t \to 0$ becomes

$$\frac{\partial w^i}{\partial x^k}v^k - w^j\frac{\partial v^i}{\partial x^j} = [v,w]x^i .$$

Definition (A.2) also provides formulas for tensors of other types, but given the Lie derivative on scalars and vector fields, such formulas can more easily be derived by the Leibniz rule.

AA.2 Tensor densities

For a spatial metric h_{ab}, its determinant is a function but *not* a scalar because its values change under coordinate transformations. The combination $\sqrt{\det h}\,\mathrm{d}^3x$ behaves tensorially, but not $\sqrt{\det h}$ alone. This is captured by introducing a new type of covariant object, *densities* $\pi^{a_1\ldots a_k}{}_{b_1\ldots b_l}$, which can be thought of as being obtained from tensors by multiplying with a power of the determinant of the metric. This definition is independent of the dimension and signature of the Riemannian manifold, i.e. it can be used for objects on space or on space-time, or on any manifold of dimension $D \in \mathbb{N}$ and any signature.

A *tensor density* $\pi^{a_1\ldots a_k}{}_{b_1\ldots b_l}$ of *weight* $n \in \mathbb{R}$ is an object defined on a differentiable manifold by a collection of components transforming under changes of coordinates $x^a \mapsto x'^{a'}$ by

$$\pi'^{a'_1\ldots a'_k}{}_{b'_1\ldots b'_l} = \left|\det\left(\frac{\partial x^c}{\partial x'^{c'}}\right)\right|^n \pi^{a_1\ldots a_k}{}_{b_1\ldots b_l}\frac{\partial x'^{a'_1}}{\partial x^{a_1}}\cdots\frac{\partial x'^{a'_k}}{\partial x^{a_k}}\frac{\partial x^{b_1}}{\partial x'^{b'_1}}\cdots\frac{\partial x^{b_l}}{\partial x'^{b'_l}}\,.$$

In particular, a tensor density of weight zero is a tensor, and $\sqrt{|\det g|}$ a scalar density of weight one. It follows immediately that the density weight is additive under tensor multiplication. Any tensor density of weight n can be dedensitized by multiplying it with $|\det g|^{-n/2}$.

Further examples of tensor densities are the objects $\epsilon^{a_1\cdots a_D}$, defined such that they are antisymmetric in all indices and $\epsilon^{1\ldots D} = 1$ *in any coordinate system*, and $\bar{\epsilon}_{a_1\ldots a_D}$, defined such that it is antisymmetric in all indices and $\bar{\epsilon}_{1\ldots D} = 1$ *in any coordinate system*. (This appears to be the same definition in both cases, except that indices take different positions.) As shown in Exercise A.4, $\epsilon^{a_1\cdots a_D}$ has density weight one, and $\bar{\epsilon}_{a_1\ldots a_D}$ density weight -1; they are thus two different tensor densities. In particular, $\bar{\epsilon}_{a_1\ldots a_D}$ cannot be obtained by lowering the indices of $\epsilon^{a_1\cdots a_D}$, as this operation would not change the density weight.

The tensor densities $\epsilon^{a_1\cdots a_D}$ and $\bar{\epsilon}_{a_1\ldots a_D}$ take the same constant values in any coordinate system and can thus reasonably be defined to have vanishing covariant derivative: $\nabla_a\epsilon^{a_1\cdots a_D} = 0 = \nabla_a\bar{\epsilon}_{a_1\ldots a_D}$. Observing (3.39), this implies that $\nabla_a \det g = 0$, since the metric g_{ab} is also covariantly constant.

For a tensor density $\pi^{a_1\cdots a_k}{}_{b_1\ldots b_l}$ of weight n, $|\det g|^{-n/2}\,\pi^{a_1\cdots a_k}{}_{b_1\ldots b_l}$ is a tensor without density weight, for which we can use the standard formula for its covariant derivative. Using the Leibniz rule, this implies that

$$\begin{aligned}\nabla_a\pi^{a_1\ldots a_k}{}_{b_1\ldots b_l} &= |\det g|^{n/2}\,\nabla_a\left(|\det g|^{-n/2}\,\pi^{a_1\ldots a_k}{}_{b_1\ldots b_l}\right)\\ &= \partial_a\pi^{a_1\ldots a_k}{}_{b_1\ldots b_l} + \Gamma^{a_1}_{ac}\pi^{ca_2\ldots a_k}{}_{b_1\ldots b_l} + \cdots + \Gamma^{a_k}_{ac}\pi^{a_1\ldots a_{k-1}c}{}_{b_1\ldots b_l}\\ &\quad -\Gamma^c_{ab_1}\pi^{a_1\ldots a_k}{}_{cb_2\ldots b_l} - \cdots - \Gamma^c_{ab_l}\pi^{a_1\ldots a_k}{}_{b_1\ldots b_{l-1}c} - n\Gamma^b_{ba}\pi^{a_1\ldots a_k}{}_{b_1\ldots b_l}\end{aligned}$$

for a tensor density $\pi^{a_1\cdots a_k}{}_{b_1\ldots b_l}$ of weight n. Notice the last term, which arises from the partial derivative $\partial_a \log\det g = 2\Gamma^b_{ba}$.

Also, the Lie derivatives of $\epsilon^{a_1\cdots a_D}$ and $\bar{\epsilon}_{a_1\ldots a_D}$ are zero, but $\mathcal{L}_v g_{ab}$ is in general non-zero. With

$$\mathcal{L}_v \det g = 2\det g\nabla_a v^a$$

(using again (3.39) and (3.40) implied by it) the Lie derivative of a tensor density $\pi^{a_1\cdots a_k}{}_{b_1\ldots b_l}$ of weight n is

$$\mathcal{L}_v\pi^{a_1\ldots a_k}{}_{b_1\ldots b_l} = |\det g|^{n/2}\,\mathcal{L}_v\left(|\det g|^{-n/2}\,\pi^{a_1\ldots a_k}{}_{b_1\ldots b_l}\right) + n\pi^{a_1\ldots a_k}{}_{b_1\ldots b_l}\nabla_a v^a \quad \text{(A.3)}$$

where the right-hand side requires only the Lie derivative of a tensor without density weight.

Tensor densities arise naturally in canonical formulations of field theories where the space-time metric is considered one of the physical fields. In a Liouville term of the form $\int \mathrm{d}^3x\dot{\varphi}p_\varphi$ as it is required for a Legendre transformation, one cannot insert the square root of the determinant of the metric as a measure factor, since this would make φ and p_φ non-canonical. For scalars φ and p_φ, on the other hand, the simple coordinate measure d^3x would not be appropriate for coordinate-independent integrations. However, if p_ϕ has density weight one, the product d^3xp_φ transforms invariantly. Similarly, the variable conjugate to a 1-form A_a must be a vector field E^a of density weight one.

Density weights of canonical variables influence the appearance of the diffeomorphism constraint. It must generate Lie derivatives of the canonical variables, and those expressions depend on the density weight; the variables in a canonical pair thus appear in different forms in the constraint. For a scalar φ with its momentum p_φ, the constraint is $D[N^a] = \int \mathrm{d}^3x N^a p_\varphi \partial_a\varphi$, generating different transformations for φ and p_φ according to their density weights.

AA.3 Geometry of Lie groups and Lie algebras

A *Lie group* G is a smooth manifold whose elements form a group with smooth composition. There are thus diffeomorphisms $\cdot^{-1}\colon h \mapsto h^{-1}$, $L_g\colon h \mapsto gh$ and $R_g\colon h \mapsto hg$ for all $g \in G$. Left-translation L_g and right-translation R_g define actions of the group on itself. These actions are transitive: for any pair $(h_1, h_2) \in G \times G$ there is a $g \in G$ such that $h_2 = L_g h_1$; and free: g is unique, $g = h_2^{-1}h_1$. (The free, transitive group action has been used in Chapter 4.1 to construct and classify homogeneous cosmological models without isotropies.)

Differential geometry can be performed on a Lie group, introducing various kinds of vector and tensor field. Of particular importance are left-invariant vector fields X satisfying $L_{g*}X = X$. The Lie commutator of a pair of left-invariant vector fields is again left-invariant, using $\Phi_*[X, Y] = [\Phi_*X, \Phi_*Y]$ for a diffeomorphism Φ. Left-invariant vector fields thus form a *Lie algebra*, a linear space equipped with a bracket operation $[\cdot, \cdot]$ defined on it. When a Lie algebra is defined as the set of invariant vector fields of a Lie group G, we call it $\mathcal{L}G$. Choosing a basis X_I of the Lie algebra, we define the *structure constants* via $[X_J, X_K] = C^I_{JK}X_I$.

Transitivity of L_g means that a left-invariant vector field X is completely determined by the value it takes at one point, conveniently chosen as the identity element $h = \mathbb{I}$ of the Lie group. In particular, we can identify the Lie algebra $\mathcal{L}G$ with the tangent space of G at the identity, and thus $\dim \mathcal{L}G = \dim G$. A basis of the tangent space at the identity provides a set of generators T_I of the Lie algebra. As another consequence, every Lie group possesses a global tangent-space basis obtained by left-translating the tangent basis at $\mathbb{I}$; every Lie group is parallelizable. A 2-sphere cannot be the manifold of a Lie group, while the 3-sphere is realized as the group manifold of SU(2).

For a matrix group, a Lie group realized as a subgroup of the general linear group GL(n) of invertible $n \times n$-matrices, we can use matrix elements x^α_β as global coordinates (transported from a neighborhood of the identity to all of G by left translation). In these coordinates, we expand a vector in the tangent space at unity as $X|_{\mathbb{I}} = X^\alpha_\beta(\partial/\partial x^\alpha_\beta)|_{\mathbb{I}}$. Transporting it to all of G, we obtain the corresponding left-invariant vector field

$$X|_g = (L_{g*}X)|_g = \left(\frac{\partial (L_{g*}(\mathbb{I}))^\alpha_\beta}{\partial x^\gamma_\delta} X^\gamma_\delta\right) \frac{\partial}{\partial x^\alpha_\beta}\bigg|_g = \frac{\partial (x(g)^\alpha_\epsilon x^\epsilon_\beta)}{\partial x^\gamma_\delta} X^\gamma_\delta \; \frac{\partial}{\partial x^\alpha_\beta}\bigg|_g$$
$$= x(g)^\alpha_\gamma X^\gamma_\beta \; \frac{\partial}{\partial x^\alpha_\beta}\bigg|_g = \mathrm{tr}\left((gX)^T \frac{\partial}{\partial g}\right)$$

evaluated at a fixed but arbitrary $g \in G$ represented by coordinates $x(g)^\alpha_\beta$.

Integration is used to invert the transition from a Lie group to its Lie algebra. From an element X of the Lie algebra, we obtain a left-invariant vector field $X(g)$ on G, whose integral curves $h(t) \subset G$ are found by integrating $(\mathrm{d}h/\mathrm{d}t)|_g = X(g)$. For a fixed X, this defines a flow on G. Starting at the unit element and computing the flows of all $X \in \mathcal{L}G$ up to $t = 1$ produces a Lie group, which can be shown to be the universal covering of the Lie group from which $\mathcal{L}G$ was obtained. (By the integration procedure, a simply-connected manifold is obtained.) Since X is constant along its integral curves, obtained by left translation, the flow equations for a matrix group can be solved by the matrix exponential

$$\exp(tX) = \sum_{n=0}^{\infty} \frac{1}{n!}(tX)^n \,. \tag{A.4}$$

The *exponential map* $\exp\colon \mathcal{L}G \to G$ thus provides a mapping from the Lie algebra to a Lie group, inverting the descent from a Lie group to its Lie algebra as the tangent space at the unit element.

By duality with invariant vector fields, we obtain invariant 1-forms $\boldsymbol{\omega}^I$ such that $\boldsymbol{\omega}^I(X_J) = \delta^I_J$ and $L^*_g\boldsymbol{\omega}^I = \boldsymbol{\omega}^I$ for all $g \in G$. By the same calculation as used in Chapter 4.1, the structure constants in the Lie algebra of left-invariant vector fields appear in the *Maurer–Cartan relations* $\mathrm{d}\boldsymbol{\omega} = -\frac{1}{2}[\widehat{\boldsymbol{\omega},\boldsymbol{\omega}}]$, or

$$\mathrm{d}\boldsymbol{\omega}^I = -\frac{1}{2} C^I_{JK} \boldsymbol{\omega}^J \wedge \boldsymbol{\omega}^K \,. \tag{A.5}$$

With a basis of left-invariant 1-forms $\boldsymbol{\omega}^I$ dual to left-invariant vector fields associated with generators T_I of the Lie algebra, we define the *Maurer–Cartan form* $\boldsymbol{\theta}_{\mathrm{MC}} := \boldsymbol{\omega}^I T_I$, a left-invariant 1-form on G taking values in $\mathcal{L}G$. It follows that $\boldsymbol{\theta}_{\mathrm{MC}}(v^J X_J) = v^J T_J$ for arbitrary coefficients v^J (a vector at the tangent space of unity). Thus, the Maurer–Cartan form restricted to left-invariant vector fields representing the Lie algebra, $\boldsymbol{\theta}_{\mathrm{MC}}|_{\mathcal{L}G} = \mathrm{id}$, acts as the identity map. While it is invariant under left translations, it satisfies $R^*_g\boldsymbol{\theta}_{\mathrm{MC}} = \mathrm{Ad}_{g^{-1}}\boldsymbol{\theta}_{\mathrm{MC}}$ under right translations. Both properties also follow from the formula derived next.

In coordinates for a matrix group, the dual-basis relationships

$$\mathrm{d}x^\alpha_\beta \left(\frac{\partial}{\partial x^\gamma_\delta}\right) = \delta^\alpha_\gamma \delta^\delta_\beta = \mathrm{id}^{\alpha\delta}_{\gamma\beta}$$

show that the Maurer–Cartan form at the unit element of G is simply $(\boldsymbol{\theta}_{\mathrm{MC}}|_{\mathbb{I}})^{\alpha}_{\beta} = \mathrm{d}x^{\alpha}_{\beta}$. With $\boldsymbol{\omega}^I|_g = L^*_{g^{-1}}\boldsymbol{\omega}^I|_{\mathbb{I}}$ for left-invariant 1-forms, we thus obtain the expression

$$(\boldsymbol{\theta}_{\mathrm{MC}}|_g)^{\alpha}_{\beta} = L^*_{g^{-1}}\mathrm{d}x^{\alpha}_{\beta} = \mathrm{d}(x(g^{-1}))^{\alpha}_{\beta} = \frac{\partial(x(g^{-1})^{\alpha}_{\gamma}x^{\gamma}_{\beta})}{\partial x^{\delta}_{\epsilon}}\mathrm{d}x^{\delta}_{\epsilon}|_g$$
$$= x(g^{-1})^{\alpha}_{\delta}\mathrm{d}x^{\delta}_{\beta}|_g = (g^{-1}\mathrm{d}g)^{\alpha}_{\beta}\,.$$

This provides a simple way to compute the Maurer–Cartan form, or a basis of left-invariant 1-forms, for matrix groups.

AA.4 Fiber bundles, connections and *n*-beins

Several first-order formulations of general relativity exist whose basic fields are not the metric tensor but rather, connections and tetrads (on space-time) or triads (on space). Such formulations are often related to additional geometrical structures that do not necessarily arise in metric formulations.

AA.4.1 Fiber bundles

Many field theories are described by space-time dependent objects taking values in a vector space V. The basic mathematical structure behind this notion is that of *vector bundles*, which can be understood as a base manifold M (such as space-time) with a copy of a vector space V attached to each point x. In this way, a generalization of the tangent space arises with internal spaces V_x, called fibers of the vector bundle, not associated with any directions in the base manifold M. Like the tangent bundle, a vector bundle in general is not trivial, i.e. not of the form $M \times V$. A physical field is then a section of the vector bundle, which assigns to each point $x \in M$ a vector $v^A \in V_x$. Examples are the gauge theories of particle physics, whose fibers are representation spaces of the gauge groups, and gravity which can be formulated for tangent-space tensor fields or for internal vector spaces in n-bein formulations.

A *fiber bundle* (B, M, π) over M with fiber F is a differentiable manifold B together with a smooth surjective map $\pi: B \to M$ with pre-image $\pi^{-1}(x) \cong F$ for all $x \in M$, such that the fibration is locally trivial: for an open covering $M = \bigcup_{\lambda} \mathcal{U}_{\lambda}$ of the base manifold M, $\pi^{-1}(\mathcal{U}_{\lambda}) \cong \mathcal{U}_{\lambda} \times F$ for all λ. There is thus a family of diffeomorphisms f_{λ} such that every $p \in \pi^{-1}(\mathcal{U}_{\lambda})$ is smoothly identified with $(\pi(p), f_{\lambda}(p))$. Restricting the map to $\pi^{-1}(x)$ provides an identification $f_{\lambda}|_x: \pi^{-1}(x) \to F$ of all fibers with the general F. However, the identification in general depends on the neighborhood used, and thus different overlapping neighborhoods can give rise to different identifications of fiber elements. An illustration is given in Fig. 4.3 on page 133.

To take this into account, we consider the transition functions $g_{\lambda\mu}(x) := f_{\lambda}|_x \circ (f_{\mu}|_x)^{-1}: F \to F$ defined for x in the overlap region $\mathcal{U}_{\lambda} \cap \mathcal{U}_{\mu}$. These maps show how the identification of points on a fiber changes when a different local trivialization is used. All transition functions are invertible, with $g^{-1}_{\lambda\mu} = g_{\mu\lambda}$, and can be composed with each other should at least three neighborhoods overlap at x, in which case $g_{\lambda\mu} \circ g_{\mu\nu} = g_{\lambda\nu}$. The transition functions form a groupoid: a set of invertible elements $g_{\mu\nu}$ with two maps $s(g_{\mu\nu}) := \mathcal{U}_{\mu}$ and $t(g_{\mu\nu}) := \mathcal{U}_{\nu}$ in which composition of two elements $g^{(1)}$ and $g^{(2)}$ is not always defined but is so if $t(g^{(1)}) = s(g^{(2)})$. In most cases, one is interested in *G-bundles* for which the transition functions take values in a representation of some structure group G acting on F, for instance the general linear group acting on a vector space F.

A *principal fiber bundle* (P, G, M, π) is a G-bundle (P, M, π) with fibers $F = G$ and G acting on itself by the left translation $L_g\colon G \to G, h \mapsto g \cdot h$. An important property of a principal fiber bundle is that it carries a natural *right action* of the structure group G on all of P, defined by $R_g p = f_\lambda|_{\pi(p)}^{-1}(f_\lambda|_{\pi(p)} \cdot g)$. In the definition, we refer to neighborhoods and the fiber identifications they provide by local trivializations, but the right action of G is independent of that extra choice. Indeed,

$$\begin{aligned} f_\lambda|_{\pi(p)} \cdot g &= \left(g_{\lambda\mu}(\pi(p)) f_\mu|_{\pi(p)}\right) \cdot g = g_{\lambda\mu}(\pi(p)) \left(f_\mu|_{\pi(p)} \cdot g\right) \\ &= f_\lambda|_{\pi(p)} \circ f_\mu|_{\pi(p)}^{-1} \left(f_\mu|_{\pi(p)} \cdot g\right) . \end{aligned}$$

Applying $f_\lambda|_{\pi(p)}^{-1}$ to both ends of this equation shows that R_g is independent of the choice of local trivialization.

The right action allows the identification of left-invariant *vertical vector fields*, vector fields v^a on P satisfying $\pi_* v^a = 0$, with the Lie algebra of the structure group. For every $X \in \mathcal{L}G$ there is a vertical vector field $\tilde{X}$ on P for which $\tilde{X}|_p = \mathrm{d}R_{\exp(tX)} p/\mathrm{d}t$.

Physically, the right action by the structure group of a principal fiber bundle encodes gauge transformations, with gauge fields realized by connections as described in the next section. Other fields subject to gauge transformations are formulated by means of *associated vector bundles* over M with the same structure group and a vector space F as fiber. One can define an associated vector bundle with fiber F as one having the same transition functions $g_{\lambda\mu}$ as the principal fiber bundle it is associated with. More compactly, the associated fiber bundle is the manifold $P \times_\rho F := (P \times F)/\sim$, defined as the Cartesian product of P and F, factored out by the equivalence relation $(R_g p, f) \sim (p, \rho(g) f)$ where ρ introduces the action of the structure group on the fiber. Or, one can realize the associated fiber bundle as $(P \times F)/G$, the Cartesian product modulo the group action $G\colon (p, f) \mapsto (R_{g^{-1}} p, \rho(g) f)$ of G.

The fact that G acts transitively on the fibers of P means that $(P \times F)/G$ is locally $M \times F$ (while $P \times F$ is locally $M \times G \times F$). Thus, the associated bundle is a G-bundle over M with fiber F. In an overlap region $\mathcal{U}_\lambda \cap \mathcal{U}_\mu$ of M, a point $p \in P$ is decomposed as $(\pi(p), h_\lambda(p)) = (\pi(p), g_{\lambda\mu}(\pi(p)) h_\mu(p))$ with the transition functions $g_{\lambda\mu}$ of P. Thus, on $(P \times F)/\sim$ we have

$$\begin{aligned} &(\pi(p), h_\lambda(p), f) \sim (\pi(p), \mathbb{I}, \rho(h_\lambda(p)) f) = (\pi(p), \mathbb{I}, f_\lambda(p)) \\ &\qquad \sim (\pi(p), \mathbb{I}, \rho(g_{\lambda\mu}(\pi(p)) h_\mu(p)) f) = (\pi(p), \mathbb{I}, \rho(g_{\lambda,\mu}(\pi(p))) f_\mu(p)) \end{aligned}$$

confirming that the transition functions are the same for P and the associated bundle.

AA.4.2 Connections

To define derivatives of vector-bundle fields, one needs a prescription to compare vectors at different points. As on the tangent space, this is done by introducing a *connection 1-form* $\Gamma_{a\ C}^{B}$, an object such that $t^a \Gamma_{aC}^B$ evaluated at a point x is a linear map $v^B \mapsto t^a \Gamma_{aC}^B v^C$ on the vector space V_x for every vector t^a tangent to M. For a curve in M, this allows us to define

$$\delta v^B = -t^a \Gamma_{aC}^B v^C \delta t \tag{A.6}$$

as the infinitesimal displacement of an initial vector v^C at a point x along the curve with tangent vector t^a at x. If the vector spaces V_x are representation spaces of a Lie group, allowed values of $v^a\Gamma^B_{a\,C}$ are usually restricted to be in the Lie algebra of the group. In this case, the connection components are often written as Γ^i_a such that $\Gamma^B_{aC} = \Gamma^i_a(T_i)^B{}_C$ for generators T_i of the Lie algebra. For the theory of weak interactions, for instance, the group is SU(2) with three generators given by the Pauli matrices. The connection components correspond to the gauge fields Γ^i_a, $i = 1, 2, 3$, given by the three vector bosons Z and $W^{\pm}$.

Integrating the infinitesimal displacements (A.6) along the curve then provides an expression in the Lie group, which is the parallel transport or holonomy from one vector space to another one. Using the form in terms of the generators provides compact expressions as matrix exponentials, where, however, the non-Abelian case requires care with the factor ordering. In this form, integrations of the infinitesimal transport equation are usually written as

$$h_e(\Gamma) = \mathcal{P}\exp(\int_e t^a\Gamma^i_a T_i \mathrm{d}\lambda) \tag{A.7}$$

where $e(\lambda)$ is the curve involved with tangent vector t^a. This expression represents a solution to the parallel-transport equation

$$\frac{\mathrm{d}h_e(\Gamma)}{\mathrm{d}\lambda} = \Gamma^i_a t^a T_i h_e(\Gamma) \tag{A.8}$$

with a λ-dependence of $h_e(\Gamma)$ by the endpoint of the curve $e(\lambda)$ taken from 0 to λ.

Vector fields in gauge theories are sections of associated bundles. Instead of working with these sections, a connection can equivalently be introduced as a covariant derivative on the whole bundle, or on the principal fiber bundle with which it is associated. A *connection* on the principal fiber bundle P is defined as a 1-form $\boldsymbol{\theta}$ on P taking values in the Lie algebra of the structure group, such that $R^*_g\boldsymbol{\theta} = \mathrm{Ad}_{g^{-1}}\boldsymbol{\theta}$ for all $g \in G$ and $\boldsymbol{\theta}(\tilde{X}) = X$ for all $X \in \mathcal{L}G$ ($\tilde{X}$ being the vertical vector field introduced before). These two properties show that a connection generalizes the notion of the Maurer–Cartan form from a Lie group G to a principal fiber bundle with structure group G.

Unlike the Maurer–Cartan form, a connection is not uniquely defined by the requirements in its definition. It is determined when evaluated on vertical vector fields by the relationship to the Lie algebra of G, but free on the part transversal to the fibers. Indeed, a connection amounts to specifying what the horizontal direction at every point $p \in P$ should mean, which, unlike the vertical direction along the fibers, is not determined by the structure of a principal fiber bundle. With a connection $\boldsymbol{\theta}$, the horizontal space $H \subset TP$ is defined as the kernel of $\boldsymbol{\theta}$: $\boldsymbol{\theta}(H) = 0$. (The kernel cannot have a vertical component, owing to the transitivity of the right action of G on P.)

With a notion of horizontality given by a connection $\boldsymbol{\theta}$, we can lift tangent vectors $v \in TM$ to horizontal vectors $\tilde{v} \in TP$ such that $\boldsymbol{\theta}(\tilde{v}) = 0$ and $\pi_*\tilde{v} = v$. Through any point in the fiber over x, there is a unique lift of a given vector at x. Given a starting point in the fiber, we can even lift a whole curve e from a point in $\pi^{-1}(e(0))$ to one in $\pi^{-1}(e(1))$ such that it is everywhere horizontal. *Parallel transport* is defined as the map from $\pi^{-1}(e(0))$ to $\pi^{-1}(e(1))$ obtained by lifting the curve e to all starting points in the fiber over $e(0)$.

The relationship between connections as 1-forms on a principal fiber bundle and as coefficients in a covariant derivative of sections can be seen using local connection 1-forms. Given a local trivialization of P, we split elements (x, h) of P into base points x and fiber points h, and a vector field (v_M, v_G) into its base and fiber components. Since the evaluation of a connection on vertical vector fields v_G is fixed, we decompose it as

$$\boldsymbol{\theta}_{(x,h)}(v_M, v_G) = \mathrm{Ad}_{h^{-1}}\mathbf{A}_x(v_M) + \boldsymbol{\theta}_{\mathrm{MC}}|_h(v_G)\,. \qquad \text{(A.9)}$$

The local connection 1-forms $\mathbf{A}$ are defined only with respect to a local trivialization; they change by gauge transformations

$$\mathbf{A}^{(\mu)} = \mathrm{Ad}_{g_{\lambda\mu}^{-1}}\mathbf{A}^{(\lambda)} + \mathrm{Ad}_{g_{\lambda\mu}^{-1}}\boldsymbol{\theta}_{\mathrm{MC}} \qquad \text{(A.10)}$$

when the local trivialization is changed by $h_\lambda \mapsto g_{\lambda\mu}h_\mu$.

AA.4.3 n-beins

On a vector bundle, we have two different kinds of vector and tensor fields, those taking values in the fibers and those taking values in the tangent space to the base manifold. Relating these objects requires an additional structure, which is usually introduced in the context of fibers equipped with a metric η_{AB}.

An *n-bein* e^a_A, $A = 1, \ldots, n$ in an n-dimensional Riemannian manifold (a *tetrad* in four dimensions, a *triad* in three dimensions, a *dyad* in two dimensions) is an orthonormal basis of vector fields, such that $e^a_A e_{Ba} = \eta_{AB}$. One can view the label A simply as an index to enumerate the independent vector fields in the orthonormal basis of TM, or more usefully take the whole object e^a_A as a section of a fiber bundle obtained as the tensor product of the tangent bundle and another vector bundle over M. The symbol η_{AB} in the normalization condition of n-beins is then indeed the metric on the fibers.

We can also contract tetrads on the internal indices, for which we obtain $\eta^{AB}e^a_A e_{Bb} = \delta^a_b$ from the line

$$\eta^{AB}e^a_A e_{Bb}e^b_C = \eta^{AB}e^a_A\eta_{BC} = e^a_C = \delta^a_b e^b_C$$

and the fact the e^b_C as a matrix must be invertible for $\{e^a_A\}_{A=1,\ldots,n}$ to form an orthonormal basis. The inverse matrix of e^a_A differs notationally only by index positions, since from the previous equations it follows that it is given by $e^A_a = \eta^{AB}e^b_B g_{ab}$. This shows another relationship between an n-bein and the space-time metric: $e^A_a e_{bA} = g_{ab}$. Thus, an n-bein can replace the space-time metric, which then becomes a derived concept.

We are now dealing with a vector bundle over space-time M equipped with two normed vector spaces attached to each point x. There is the usual tangent space T_xM with metric $g_{ab}(x)$, but also the copy V_x of the internal vector space with the constant metric η_{AB}. (From the point of view of transformations on space-time, this internal metric is a scalar, in agreement with the fact that it does not have space-time indices.) An n-bein provides pointwise isometries between these spaces: the map $e^a_A(x)\colon V_x \to T_pM$, $v^A \mapsto e^a_A v^A$ and the inverse $e^A_a(x)\colon T_xM \to V_x$. In more methodological terms, n-beins can be used to replace all space-time indices $a, \ldots$ by internal indices $A, \ldots$ and vice versa.

Just as a metric defines a unique compatible connection, the Christoffel connection on the tangent bundle, an n-bein defines a unique compatible connection. Its connection 1-forms can be computed by $\omega_{aAB} = e^b_A\nabla_a e_{Bb}$, where ∇_a is the derivative operator compatible with the metric defined by the tetrad. With the connection 1-forms ω_{aAB}, a covariant derivative and associated parallel transport is defined for internal vector and tensor fields, such as $\mathcal{D}_a v^A = \nabla_a v^A + \omega_a{}^A{}_B v^B$. Applying a covariant derivative to the internal metric η_{AB}, which is constant, implies that

$$\mathcal{D}_a\eta_{AB} = \nabla_a\eta_{AB} - \omega_a{}^C{}_A\eta_{CB} - \omega_a{}^C{}_B\eta_{AC} = -\omega_{aBA} - \omega_{aAB} = 0\,.$$

Thus, the connection 1-forms must be antisymmetric in the internal indices, in agreement with the fact that $t^a \omega_a{}^A{}_B$, according to the general concepts for connections, must take values in the Lie algebra of the group preserving η_{AB} (for instance, the Lorentz group for an internal Minkowski space). In terms of a basis T_i generating the Lie algebra, we define the spin connection coefficients ω_a^i via $\omega_a{}^A{}_B = \omega_a^i (T_i)^A{}_B$.

Any n-bein determines a unique spin connection for which it is covariantly constant: $\mathcal{D}_a e_b^A = 0$. Writing out the covariant derivative with connection components, we have $\partial_a e_b^B - \Gamma_{ab}^c e_c^B + \Gamma_{aA}^B e_b^A = 0$ with the Christoffel connection Γ_{ab}^c, and thus $\Gamma_{aA}^B = -e_A^b(\partial_a e_b^B - \Gamma_{ab}^c e_c^B)$. Using the general expression of the Christoffel connection in terms of the metric, which in turn is related to the n-bein, we write

$$\eta_{FG}\Gamma_{aE}^G = e_E^b e_{cF}\Gamma_{ab}^c - e_E^b \partial_a e_{bF} = 2(\partial_{[b} e_{a][F}) e_{E]}^b + e_a^B e_{[E}^b e_{F]}^c \partial_b e_c^B \,. \tag{A.11}$$

In special dimensions, this can be written in alternative ways; see Exercise A.8.

AA.5 Poisson geometry

In contrast to the geometry of space-time — which is determined by a symmetric covariant 2-tensor, the metric g_{ab} — the geometry of phase space is determined by an antisymmetric contravariant 2-tensor, the Poisson tensor $\mathcal{P}^{ij}$. In addition to the antisymmetry, it must satisfy the Jacobi identity

$$\mathcal{P}^{k[i}\partial_k \mathcal{P}^{jl]} = 0 \tag{A.12}$$

This tensor provides the Poisson bracket $\{f, g\} := \mathcal{P}^{ij}(\partial_i f)(\partial_j g)$ for any pair of phase-space functions f and g. A Poisson tensor also provides a unique association of a *Hamiltonian vector field* generated by a phase-space function H: the vector field $\mathcal{P}^{ij}(\partial_j H)\partial_i$ in a coordinate basis, coordinate independently also denoted as $\mathcal{P}^\sharp \mathrm{d}H$ where $\mathcal{P}^\sharp \colon T^*M \to TM$ is the map obtained by lifting co-tangent space indices using the Poisson tensor.

If $\mathcal{P}^{ij}$ is invertible, the phase space is a symplectic manifold with symplectic form $\Omega_{ij} = (\mathcal{P}^{ij})^{-1}$. The Jacobi identity then implies that Ω_{ij} is indeed a closed 2-form, as required for a symplectic form. In this case, the phase space is also endowed with symplectic geometry.

AA.5.1 Local structure of Poisson manifolds

A Poisson manifold can be thought of as built from *symplectic leaves*, which provide a (possibly singular) foliation of the manifold. Symplectic leaves are defined as the integral surfaces of the distribution $\mathcal{P}_x^\sharp(T_x^* M) \subset T_x M$, which is guaranteed to be integrable by the Jacobi identity. The co-normal space to the foliation, given by ker $\mathcal{P}^\sharp$, is locally given by *Casimir functions* C^I such that $\mathcal{P}^\sharp(\mathrm{d}C^I) = (\mathcal{P}^{ij}\partial_i C^I)\partial_j = 0$ is the zero vector field. Factoring out the co-normal directions removes the kernel from the Poisson tensor. On the leaves a symplectic structure is induced.

Indeed, locally one can always choose *Casimir–Darboux coordinates* (x^α, C^I) such that the Poisson tensor is

$$(\mathcal{P}^{ij}) = \begin{pmatrix} \Pi^{\alpha\beta} & 0 \\ 0 & 0 \end{pmatrix} \tag{A.13}$$

with

$$(\Pi^{\alpha\beta}) = \begin{pmatrix} 0 & 1 & & 0 & 0 \\ -1 & 0 & & 0 & 0 \\ & & \ddots & & \\ 0 & 0 & & 0 & 1 \\ 0 & 0 & & -1 & 0 \end{pmatrix}$$

the standard Poisson tensor of a symplectic manifold in Darboux coordinates. The constant $\Pi^{\alpha\beta}$ can be inverted to give the leaf-symplectic structure $\Omega_{\alpha\beta}$.

Example A.1 *Let $M = g^*$ be the dual of a Lie algebra g with structure constants $C^{ij}{}_k$. A Poisson structure, called the Kirillov–Kostant structure, is naturally defined on M by $\mathcal{P}_X(\alpha, \beta) = X([\alpha, \beta]_g)$ at $X \in g^*$. This defines the Poisson tensor at any point $X \in g^*$, applied to a pair of 1-forms $\alpha, \beta \in T^*_X g^* \cong g^{**} = g$. By the latter identification of g^{**} with g, the Lie algebra bracket on g can directly be used to define the Poisson bracket between forms on g^*. In coordinates in which $X = X^k T_k$ with Lie algebra generators T_k, this implies a linear Poisson tensor $\mathcal{P}^{ij}(X) = C^{ij}{}_k X^k$. Casimir functions in the Poisson-geometry sense then agree with Casimir functions of the Lie algebra. For $g = \mathrm{su}(2)$, the symplectic leaves are 2-spheres.*

In general, symplectic forms $\Omega_{\alpha\beta}$ defined on the leaves of a Poisson manifold cannot be combined to a presymplectic form on the whole Poisson manifold. If this is possible, we define, following Bojowald and Strobl (2003a), as follows:

Definition A.1 *A* compatible presymplectic form *on a Poisson manifold $(M, \mathcal{P})$ is a closed 2-form $\tilde{\mathbf{\Omega}}$ on M such that $\iota_L^* \tilde{\mathbf{\Omega}} = \mathbf{\Omega}_L$ for all leaves L in $(M, \mathcal{P})$, where $\iota_L \colon L \to M$ is the embedding of a leaf L in M, and $\mathbf{\Omega}_L$ is the leaf symplectic structure induced from $\mathcal{P}$.*

Example A.2 *If the Poisson manifold is globally of the form $M = L \times \mathbb{R}^k$ with leaves $L \times (C^1, \ldots, C^k)$ with C^I constant, and if the leaf-symplectic structures are $\mathbf{\Omega}_L = \mathrm{d}_L \boldsymbol{\theta}_L$ for 1-forms $\boldsymbol{\theta}_L$ defined on the leaves (parameterized by Casimir coordinates C^I), then $\tilde{\mathbf{\Omega}} = \mathrm{d}\boldsymbol{\theta}$ is a compatible presymplectic form. (For $\tilde{\mathbf{\Omega}}$ we take the full differential on M, including derivatives by Casimir coordinates. Leaf-symplectic structures $\mathbf{\Omega}_L$, on the other hand, are obtained by keeping Casimir coordinates fixed to specify the leaf and taking derivatives in d_L only along leaf coordinates.)*

More generally, if $M = L \times \mathbb{R}^k$ is foliated trivially, a necessary condition for the existence of a compatible presymplectic form is $\partial_I \oint_\sigma \mathbf{\Omega}_L = 0$, where ∂_I is the transversal derivative along a Casimir coordinate, and σ an arbitrary closed 2-cycle in the leaf L. This condition rules out dual Lie algebras of compact Lie groups as possible Poisson manifolds with compatible presymplectic form. On the other hand, a compatible presymplectic form exists if all leaves have trivial second cohomology.

AA.5.2 Constraints

Poisson geometry plays a large role in the analysis of Hamiltonian and constrained systems. If we start with a symplectic manifold but impose constraints to reduce it to a submanifold, the pull-back of the symplectic form to the constraint surface, while automatically closed, need no longer be non-degenerate. This happens whenever the constraints are not purely second class, and the constraint surface is only a presymplectic manifold. In fact, as suggested by Bojowald and Strobl (2003b), one can take these considerations to arrive at a general definition of first- and second-class surfaces.

Definition A.2 *Let $(M, \mathcal{P})$ be a Poisson manifold, and $\iota\colon C \to M$ be an embedded submanifold such that $\iota_* TC \subset \mathcal{P}^\sharp(T^*M)$. (The submanifold C is thus contained in a symplectic leaf of $(M, \mathcal{P})$.) We call C*

(i) first class *if $\{0\} \neq \mathcal{P}^\sharp(T_x^{*\perp}C) \subset T_xC$ for all $x \in C$, and*
(ii) second class *if $\mathcal{P}^\sharp(T_x^{*\perp}C) \cap T_xC = \{0\}$.*

In both cases, we use the co-normal space $T_x^{*\perp}C := \{\alpha \in T_x^*M\colon \alpha(v) = 0$ for all $v \in TC_x\}$.

If the surface is locally described by vanishing constraints $C^I = 0$ (including Casimir functions), such that the co-normal space is spanned by $\mathrm{d}C^I$, all Hamiltonian vector fields X_{C^I} generated by the constraints are tangent to C if C is first class, while none of them is tangent to C for a second-class surface. The requirement that $\mathcal{P}^\sharp(T_x^{*\perp}C)$ not vanish for a first-class surface ensures that the constraints are not just Casimir functions of the Poisson manifold, and that C is indeed a proper subset of a symplectic leaf. (Otherwise, a symplectic leaf would be both first and second class.)

If $\mathcal{P} = \mathbf{\Omega}^{-1}$ is invertible, first- and second-class constraint surfaces in a symplectic manifold $(M, \mathbf{\Omega})$ are first- and second-class surfaces, respectively, in the sense defined in the context of the Dirac classification of constraints. The definition given here is more general and applies to any Poisson manifold.

For second-class constraints in a symplectic manifold $(M, \mathbf{\Omega})$, we can define the Dirac bracket via the Poisson bivector

$$\mathcal{P}_{\mathrm{D}} = \mathcal{P} + \sum_{I,J} (\{C^I, C^J\})^{-1} X_{C^I} \wedge X_{C^J} \tag{A.14}$$

with $\mathcal{P} = \mathbf{\Omega}^{-1}$. Its interpretation in the context of Poisson geometry is that $(M, \mathcal{P}_{\mathrm{D}})$ is a Poisson manifold in which the second-class surface is a symplectic leaf (with the constraint functions C^I as Casimir functions) such that $\mathbf{\Omega}$ provides a compatible (pre)symplectic form: the leaf-symplectic structure is $\mathbf{\Omega}_L = \iota_L^* \mathbf{\Omega}$ for all symplectic leaves of $(M, \mathcal{P}_{\mathrm{D}})$, embedded by ι_L. In terms of the Casimir functions, we realize each leaf L by the equation $C^I = C^I_{(L)}$ assigning fixed values to the C^I.

Proof We take two functions $f, g \in C^\infty(L)$ on L and compare $\{f, g\}_L = X_g f$ (computed using the leaf-symplectic structure) with $\{F, G\}_{\mathrm{D}}$ evaluated as the Dirac bracket for two arbitrary extensions F and G of f and g, respectively: $F, G \in C^\infty(M)$ such that $\iota^* F = f$ and $\iota^* G = g$. Around a fixed L, two different extensions F and F' of f are related by $F' = F + x_I(C^I - C^I_{(L)})$ for some functions x^I. With $\{C^I, \cdot\}_{\mathrm{D}} = 0$ by construction, the value of $\{F, G\}_{\mathrm{D}}$ does not depend on the extension we choose. For simplicity, we then choose F and G such that $\iota_L^*(X_{C^I} F) = 0$ and $\iota_L^*(X_{C^I} G) = 0$ for all C^I (and thus F and G are extremized on L along transversal directions). Since no X_{C^I} is tangent to L for a second-class surface as assumed, the values of $\iota_L^* F$ and $\iota_L^* G$ are not restricted by these conditions.

Then,

$$\begin{aligned}\iota_L^*\{F, G\}_{\mathrm{D}} &= \iota_L^* \left(\mathcal{P}^{ij}(\partial_i F)(\partial_j G) + (\{C^I, C^J\})^{-1}(X_{C^I} F)(X_{C^J} G)\right) \\ &= \iota_L^* \left(\mathcal{P}^{ij}(\partial_i F)(\partial_j G)\right) = \iota_L^*(X_G F)\,.\end{aligned}$$

Since $\iota_L^*(X_G C^I) = -\iota_L^*(X_{C^I} G) = 0$, X_G is tangent to L on L, so that its restriction $X_G|_L \in \Gamma(TL)$ defines a vector field on L. Moreover, the existence of an expansion $\boldsymbol{\Omega} = \boldsymbol{\Omega}_L + \boldsymbol{\rho}_I \wedge \mathrm{d}C^I + \sigma_{IJ}\mathrm{d}C^I \wedge \mathrm{d}C^J$ in a neighborhood of L with 1-forms $\boldsymbol{\rho}_I$ and functions σ_{IJ}, combined with $\iota_L^*\mathrm{d}C^I = 0$, implies that

$$\left(\mathcal{P}_L^{\sharp-1} X_G|_L\right)_\beta = (\boldsymbol{\Omega}_L)_{\alpha\beta} X_G|_L^\alpha = \iota_L^*(\boldsymbol{\Omega}_{\alpha\beta} X_G^\alpha)$$
$$= \left(\iota_L^*(\boldsymbol{\Omega}^\sharp X_G)\right)_\beta = (\iota_L^*\mathrm{d}G)_\beta = (\mathrm{d}g)_\beta$$

and thus $\iota_L^*(X_G F) = (X_G|_L F)|_L = (\mathcal{P}_L^\sharp \mathrm{d}g) f = X_g f = \{f, g\}$, completing the proof. □

AA.6 Lie algebroids

Lie algebroids are generalizations of Lie algebras, obtained by making the structure functions depend on positions in some space. Accordingly, they can be used to interpret the symmetries behind constraint algebras with structure functions. As shown in Chapter 4.3.2.4, this has been established especially for Poisson sigma models, which contain 2-dimensional dilaton gravity models as special cases. Sophisticated algebraic structures then facilitate a deeper understanding of the solution spaces of these systems. While no comparably explicit construction is available for full gravity in $3+1$ dimensions, the hypersurface-deformation algebra is another example for Lie algebroids; see Chapter 4.3.2.6. Lie algebroids thus provide several important insights about the nature of gauge transformations.

Position dependence can be formulated elegantly in the fiber-bundle language, with the base manifold M as the space of dependence. Structure functions are then elements of $C^\infty(M)$, and symmetry generators are elements of a fiber basis equipped with a Lie bracket. The appropriate notion on which algebraic relations are to be defined is thus that of sections in the fiber bundle, on which smooth functions act by multiplication. For the Lie bracket to be a derivation, we must require a Leibniz rule to be satisfied, which in turn requires a map from the fibers of the bundle to vector fields on M. These are the crucial ingredients of the following Definition

Definition A.3 *A* Lie algebroid *is a vector bundle E over a base manifold M together with a bundle map $\rho\colon E \to TM$ (the* anchor*) and a Lie algebra structure* $(\Gamma(E), [\cdot,\cdot])$ *on its sections such that the Leibniz rule*

$$[s_1, f s_2] = f[s_1, s_2] + (\rho(s_1) f)\, s_2 \tag{A.15}$$

is satisfied for all $s_1, s_2 \in \Gamma(E)$ and $f \in C^\infty(M)$.

From the Leibniz rule, it follows that the kernel of ρ defines a Lie algebra over every point in M, called the *transversal Lie algebra*.

Example A.3 *Several well-known bundles with algebraic operations defined on them can be interpreted as Lie algebroids:*

(i) *A Lie algebra g seen as a vector bundle over a single point with anchor map $\rho = 0$ is a Lie algebroid. So is a bundle of Lie algebras over a base manifold, with anchor map $\rho = 0$.*
(ii) *The tangent bundle TM over any manifold M with $\rho = \mathrm{id}$ is a Lie algebroid.*
(iii) *Consider the co-tangent space T^*M of a Poisson manifold $(M, \mathcal{P})$. The Poisson tensor defines a map $\mathcal{P}^\sharp\colon T^*M \to TM$ via $\mathcal{P}^\sharp(\alpha) = \mathcal{P}(\alpha, \cdot)$. (In components, the 1-form $\alpha = \alpha_i \mathrm{d}X^i$ is mapped*

to the vector field $\alpha_i \mathcal{P}^{ij}\partial_j$.) The map $\rho = \mathcal{P}^\sharp$ provides the anchor map on $E = T^*M$. To introduce the Lie bracket on sections, one defines $[\mathrm{d}f, \mathrm{d}g] = \mathrm{d}\{f, g\}$ on exact 1-forms and extends this to all 1-forms using linearity and the Leibniz rule. (In components, $[\mathrm{d}X^i, \mathrm{d}X^j] = \partial_k \mathcal{P}^{ij}\mathrm{d}X^k$.)

To compare and relate different Lie algebroids, the notion of morphisms is used. Since Lie algebroids are vector bundles, we start with vector bundle morphisms between two bundles $E_1 \to M_1$ and $E_2 \to M_2$. To preserve the vector-bundle structure, a morphism must map the first base manifold to the second one, and be fiberwise linear, i.e. relate the fibers over M_1 and their images in a linear way. Thus, a vector bundle morphism ϕ is obtained from a base map $\phi_0 \colon M_1 \to M_2$ together with a section of $E_1^* \otimes \phi_0^* E_2$. The latter bundle is obtained as the tensor product of the dual bundle of E_1 (having the duals of fiber spaces of E_1 as its fibers over the same base manifold M_1) and the pull-back bundle $\phi_0^* E_2$, a vector bundle over M_1 whose fiber over $x \in M_1$ is the E_2-fiber over $\phi_0(x)$. (Here, just fibers are being pulled back, not sections. Such a pull-back bundle is independent of pull-backs of differential forms, which are sections of the co-tangent bundle. The pull-back bundle exists for any vector bundle, such as the tangent bundle even though sections in the tangent bundle cannot be pulled back.) The dual action of E_1^* on E_1, mapping a section on E_1 to a function over M_1, then makes sections of $E_1^* \otimes \phi_0^* E_2$ identical to maps from $\Gamma(E_1)$ to $\Gamma(\phi_0^* E_2)$.

To take into account also the algebraic structure of Lie algebroids, we use the following Definition

Definition A.4 *A* Lie algebroid morphism *between Lie algebroids $(E_i, M_i, \rho_i, [\cdot, \cdot]_i)$, $i = 1, 2$, is a vector-bundle morphism $\phi \colon E_1 \to E_2$ such that*

$$\rho_2 \circ \phi = (\phi_0)_* \circ \rho_1 \tag{A.16}$$

(the morphisms $\phi \colon E_1 \to E_2$ and $(\phi_0)_* \colon TM_1 \to TM_2$ commute with the anchor maps ρ_1 and ρ_2) and

$$\phi \circ s_1 = s_2 \circ \phi_0 \quad \text{and} \quad \phi \circ s_1' = s_2' \circ \phi_0 \quad \text{implies} \quad \phi \circ [s_1, s_1']_1 = [s_2, s_2']_2 \circ \phi_0 \tag{A.17}$$

for all $s_i, s_i' \in \Gamma(E_i)$.

It follows directly that Lie algebra morphisms provide Lie algebroid morphisms in the case of vanishing anchor maps. If the anchor map is the identity, diffeomorphisms on the base manifold provide Lie algebroid morphisms for the tangent bundle. A non-trivial example of Lie algebroid morphisms is given by solutions of Poisson sigma models, which, as shown in Chapter 4.3.2.4, provide Lie algebroid morphisms from the tangent bundle of the worldsheet to the Poisson target manifold.

Some Lie algebroids can be realized as certain tangent spaces in Lie groupoids, generalizing the relationship between Lie algebras and Lie groups. However, not every Lie algebroid can be integrated to a Lie group; a complete classification of the integrable cases has been given by Crainic and Fernandes (2003). A Lie groupoid is a differentiable manifold G together with two maps $s \colon G \to M$ ("source") and $t \colon G \to M$ ("target") to a manifold M, such that an associative product $g_1 g_2$ is defined for any pair $(g_1, g_2) \in G$ with $t(g_2) = s(g_1)$. Moreover, there is an identity section $\mathbb{I} \colon M \to G$ such that $\mathbb{I}(t(g))g = g = g\mathbb{I}(s(g))$, and groupoid elements are invertible in the sense that for all g, $g^{-1}g = \mathbb{I}(s(g))$ and $gg^{-1} = \mathbb{I}(s(g))$.

A Lie groupoid does not act transitively on itself by multiplication, but there is a left action defined by multiplication of $g \in G$ on $t^{-1}(s(g)) \subset G$. A left-invariant vector field on G is a section of $\ker t_* \subset TG$ invariant under left multiplication. Left-invariant vector fields can be shown to form a Lie algebroid over M, identified with the subbundle of the tangent bundle at $\mathbb{I}(M)$ normal to $T\mathbb{I}(M)$ in TG. This Lie algebroid is uniquely associated with the original Lie groupoid. More details can be found in the book by Cannas da Silva and Weinstein (1999).

Exercises

A.1 Compute the Lie derivative of a covariant 2-tensor h_{ab}, such as the spatial metric, as used to derive (3.50).

A.2 Show that the symplectic form $\boldsymbol{\Omega}$ on a $2n$-dimensional phase space has a vanishing Lie derivative under Hamiltonian flow: $\mathcal{L}_{X_H}\boldsymbol{\Omega} = 0$ for $\boldsymbol{\Omega}(X_H, \cdot) = \mathrm{d}H$ with a Hamiltonian function H.

A.3 Prove Cartan's formula

$$\mathrm{d}\boldsymbol{\alpha}(v, w) = \mathcal{L}_v(\boldsymbol{\alpha}(w)) - \mathcal{L}_w(\boldsymbol{\alpha}(v)) - \boldsymbol{\alpha}(\mathcal{L}_v w)$$

for a 1-form $\boldsymbol{\alpha}$ and two vector fields v and w.

A.4 (i) Show that the square root $\sqrt{|\det g|}$ of the determinant of the metric is a density of weight one, using a general formula for the determinant.
(ii) Show that $\varepsilon^{a_1 \ldots a_D}$ is a tensor density of weight 1. Use this to give an alternative proof that $\sqrt{|\det g|}$ is of weight one based on the fact that the density weight is additive under tensor multiplication of densities.
(iii) Show that $\bar{\varepsilon}_{a_1 \ldots a_D}$ is a tensor density of weight -1.

A.5 (i) Verify explicitly that the covariant derivatives of ε^{abc} and $\bar{\varepsilon}_{abc}$ (for $D = 3$) vanish.
(ii) Show that the covariant divergence $\nabla_a E^a$ of a vector density E^a of weight one does not depend on the connection coefficients and equals the divergence in coordinate derivatives.

A.6 Compute the Lie derivative of $\sqrt{|\det g|}$, and of a general tensor density π^{ab} of weight one.

A.7 Show that a scalar density on a 1-dimensional space transforms like a 1-form under orientation-preserving coordinate changes, and a vector field of density weight one like a scalar.

A.8 Derive (A.11). On a 3-dimensional space, show that the spin connection can be expressed via $\Gamma_a^A = \frac{1}{2}\epsilon^{AB}{}_C \Gamma_{aB}^C$ as

$$\Gamma_a^A = \frac{1}{2}\epsilon^{ABC} e_C^b (2\partial_{[b} e_{a]B} + e_B^c e_{aD} \partial_b e_c^D)\,. \tag{EA.1}$$

References

Anderson, I. Fels, M. and Torre, C. (2000). Group invariant solutions without transversality and the principle of symmetric criticality. *Commun. Math. Phys.*, **212**, 653–686.

Anderson, J. L. and Bergmann, P. G. (1951). Constraints in covariant field theories. *Phys. Rev.*, **83**, 1018–1025.

Andersson, L. and Rendall, A. D. (2001). Quiescent cosmological singularities. *Commun. Math. Phys.*, **218**, 479–511.

Arnowitt, R. Deser, S. and Misner, C. W. (1962). *The Dynamics of General Relativity*. New York: Wiley. Reprinted in Arnowitt et al. (2008).

Arnowitt, R. Deser, S. and Misner, C. W. (2008). The dynamics of general relativity. *Gen. Rel. Grav.*, **40**, 1997–2027.

Ashtekar, A. (1981). Radiative degrees of freedom of the gravitational field in exact general relativity. *J. Math. Phys.*, **22**, 2885–2895.

Ashtekar, A. (1987). New Hamiltonian formulation of general relativity. *Phys. Rev. D*, **36**(6), 1587–1602.

Ashtekar, A. and Bojowald, M. (2005). Black hole evaporation: A paradigm. *Class. Quantum Grav.*, **22**, 3349–3362.

Ashtekar, A. and Bojowald, M. (2006). Quantum geometry and the Schwarzschild singularity. *Class. Quantum Grav.*, **23**, 391–411.

Ashtekar, A. and Krishnan, B. (2003). Dynamical horizons and their properties. *Phys. Rev. D*, **68**, 104030.

Ashtekar, A. and Krishnan, B. (2004). Isolated and dynamical horizons and their applications. *Living Rev. Rel.*, **7**, 1–77.

Ashtekar, A. and Lewandowski, J. (1995). Projective techniques and functional integration for gauge theories. *J. Math. Phys.*, **36**(5), 2170–2191.

Ashtekar, A. and Lewandowski, J. (1997). Quantum theory of geometry I: Area operators. *Class. Quantum Grav.*, **14**, A55–A82.

Ashtekar, A. and Lewandowski, J. (1998). Quantum theory of geometry II: Volume operators. *Adv. Theor. Math. Phys.*, **1**, 388–429.

Ashtekar, A. and Streubel, M. (1981). Symplectic geometry of radiative modes and conserved quantities at null infinity. *Proc. Roy. Soc. Lond. A*, **376**, 585–607.

Ashtekar, A. Pawlowski, T. and Singh, P. (2006). Quantum nature of the Big Bang: An analytical and numerical investigation. *Phys. Rev. D*, **73**, 124038.

Ashtekar, A. Engle, J. and Sloan, D. (2008). Asymptotics and Hamiltonians in a first order formalism. *Class. Quantum Grav.*, **25**, 095020.

Baker, J. G. Centrella, J. Choi, D. I. Koppitz, M. and van Meter, J. (2006). Gravitational-wave extraction from an inspiraling configuration of merging black holes. *Phys. Rev. Lett.*, **96**, 111102.

Balakrishna, J. Daues, G. Seidel, E. Suen, W.-M. Tobias, M. and Wang, E. (1996). Coordinate conditions in three-dimensional numerical relativity. *Class. Quantum Grav.*, **13**, L135–L142.

Barbero G., J. F. (1995). Real Ashtekar variables for Lorentzian signature space-times. *Phys. Rev. D*, **51**(10), 5507–5510.

Bardeen, J. M. (1980). Gauge-invariant cosmological perturbations. *Phys. Rev. D*, **22**, 1882–1905.

Bardeen, J. M. Carter, B. and Hawking, S. W. (1973). The four laws of black hole mechanics. *Commun. Math. Phys.*, **31**, 161–170.

Bassett, B. A. Tsujikawa, S. and Wands, D. (2006). Inflation dynamics and reheating. *Rev. Mod. Phys.*, **78**, 537–589.

Baumann, D. (2007). On the quantum origin of structure in the inflationary universe. arXiv:0710.3187.

Baumgarte, T. W. and Shapiro, S. L. (1998). Numerical integration of Einstein's field equations. *Phys. Rev. D*, **59**, 024007.

Baumgarte, T. W. and Shapiro, S. L. (2010). *Numerical Relativity: Solving Einstein's Equations on the Computer*. Cambridge University Press.

Bekenstein, J. D. (1973). Black holes and entropy. *Phys. Rev. D*, **1**, 2333–2346.

Belinskii, V. A. Khalatnikov, I. M. and Lifschitz, E. M. (1982). A general solution of the Einstein equations with a time singularity. *Adv. Phys.*, **13**, 639–667.

Ben-Dov, I. (2007). Outer trapped surfaces in Vaidya spacetimes. *Phys. Rev. D*, **75**, 064007.

Bengtsson, I. and Senovilla, J. M. M. (2009). A note on trapped surfaces in the Vaidya solution. *Phys. Rev. D*, **79**, 024027.

Berger, B. (2002). Numerical approaches to spacetime singularities. *Living Rev. Relativity*, **5**, 1. `http://www.livingreviews.org/lrr-2002-1`.

Bergmann, P. G. (1961). Observables in general relativity. *Rev. Mod. Phys.*, **33**, 510–514.

Bertschinger, E. (2006). On the growth of perturbations as a test of dark energy. *Astrophys. J.*, **648**, 797.

Bičák, J. and Hájíček, P. (2003). Canonical theory of spherically symmetric spacetimes with cross-streaming null dusts. *Phys. Rev. D*, **68**, 104016.

Bičák, J. and Kuchař, K. V. (1997). Null dust in canonical gravity. *Phys. Rev. D*, **56**, 4878–4895.

Blohmann, C. Barbosa Fernandes, M. C. and Weinstein, A. (2010). Groupoid symmetry and constraints in general relativity. 1: Kinematics. arXiv:1003.2857.

Bojowald, M. (2007). Large scale effective theory for cosmological bounces. *Phys. Rev. D*, **75**, 081301(R).

Bojowald, M. (2008). Loop quantum cosmology. *Living Rev. Relativity*, **11**, 4. `http://www.livingreviews.org/lrr-2008-4`.

Bojowald, M. (2011). *Quantum Cosmology: A Fundamental Theory of the Universe*. New York: Springer.

Bojowald, M. and Skirzewski, A. (2006). Effective equations of motion for quantum systems. *Rev. Math. Phys.*, **18**, 713–745.

Bojowald, M. and Strobl, T. (2003a). Classical solutions for Poisson sigma models on a Riemann surface. *JHEP*, **0307**, 002.

Bojowald, M. and Strobl, T. (2003b). Poisson geometry in constrained systems. *Rev. Math. Phys.*, **15**, 663–703.

Bojowald, M. and Tsobanjan, A. (2009). Effective constraints for relativistic quantum systems. *Phys. Rev. D*, **80**, 125008.

Bojowald, M. Goswami, R. Maartens, R. and Singh, P. (2005a). A black hole mass threshold from non-singular quantum gravitational collapse. *Phys. Rev. Lett.*, **95**, 091302.

Bojowald, M. Kotov, A. and Strobl, T. (2005b). Lie algebroid morphisms, Poisson sigma models, and off-shell closed gauge symmetries. *J. Geom. Phys.*, **54**, 400–426.

Bojowald, M. Cartin, D. and Khanna, G. (2007). Lattice refining loop quantum cosmology, anisotropic models and stability. *Phys. Rev. D*, **76**, 064018.

Bojowald, M. Hossain, G. Kagan, M. and Shankaranarayanan, S. (2008). Anomaly freedom in perturbative loop quantum gravity. *Phys. Rev. D*, **78**, 063547.

Bojowald, M. Sandhöfer, B. Skirzewski, A. and Tsobanjan, A. (2009a). Effective constraints for quantum systems. *Rev. Math. Phys.*, **21**, 111–154.

Bojowald, M. Hossain, G. Kagan, M. and Shankaranarayanan, S. (2009b). Gauge invariant cosmological perturbation equations with corrections from loop quantum gravity. *Phys. Rev. D*, **79**, 043505.

Bojowald, M. Reyes, J. D. and Tibrewala, R. (2009c). Non-marginal LTB-like models with inverse triad corrections from loop quantum gravity. *Phys. Rev. D*, **80**, 084002.

Bona, C. and Massó, J. (1988). Harmonic synchronizations of spacetime. *Phys. Rev. D*, **38**, 2419–2422.

Bona, C. and Massó, J. (1992). Hyperbolic evolution system for numerical relativity. *Phys. Rev. Lett.*, **68**, 1097–1099.

Bona, C. Massó, J. Seidel, E. and Stela, J. (1995). New formalism for numerical relativity. *Phys. Rev. Lett.*, **75**, 600–603.

Bondi, H. van der Burg, M. G. J. and Metzner, A. W. K. (1962). Gravitational waves in general relativity. VII. Waves from axi-symmetric isolated systems. *Proc. Roy. Soc. A*, **269**, 21–52.

Booth, I. and Fairhurst, S. (2003). Canonical phase space formulation of quasilocal general relativity. *Class. Quantum Grav.*, **20**, 4507–4532.

Booth, I. and Mann, R. (1999). Moving observers, nonorthogonal boundaries, and quasilocal energy. *Phys. Rev. D*, **59**, 064021.

Brady, P. R. Creighton, J. D. E. and Thorne, K. S. (1998). Computing the merger of black-hole binaries: The IBBH problem. *Phys. Rev. D*, **58**, 061501.

Brandt, S. and Brügmann, B. (1997). A simple construction of initial data for multiple black holes. *Phys. Rev. Lett.*, **78**, 3606–3609.

Brill, D. R. and Lindquist, R. W. (1963). Interaction energy in geometrostatics. *Phys. Rev.*, **131**, 471–476.

Brizuela, D. Martín-García, J. M. and Mena Marugán, G. A. (2007). Higher-order gauge-invariant perturbations of a spherical spacetime. *Phys. Rev. D*, **76**, 024004.

Brodbeck, O. (1996). On symmetric gauge fields for arbitrary gauge and symmetry groups. *Helv. Phys. Acta*, **69**, 321–324.

Brown, J. D. and Kuchař, K. V. (1995). Dust as a standard of space and time in canonical quantum gravity. *Phys. Rev. D*, **51**, 5600–5629.

Brown, J. D. and York, J. W. (1993). Quasilocal energy and conserved charges derived from the gravitational action. *Phys. Rev. D*, **47**, 1407–1419.

Bruni, M. Dunsby, P. K. S. and Ellis, G. F. R. (1992). Cosmological perturbations and the physical meaning of gauge invariant variables. *Astrophys. J.*, **395**, 34–53.

Campanelli, M. Lousto, C. O. Marronetti, P. and Zlochower, Y. (2006). Accurate evolutions of orbiting black-hole binaries without excision. *Phys. Rev. Lett.*, **96**, 111101.

Cannas da Silva, A. and Weinstein, A. (1999). *Geometric models for noncommutative algebras*. Berkeley Mathematics Lectures, vol. 10. Providence: Am. Math. Soc.

Cattaneo, A. S. and Felder, G. (2000). Poisson sigma models and symplectic groupoids. math/0003023.

Chinea, F. J. Fernandez-Jambrina, L. and Senovilla, J. M. M. (1992). A singularity-free space-time. *Phys. Rev. D*, **45**, 481–486.

Choptuik, M. W. (1991). Consistency of finite-difference solutions of Einstein's equations. *Phys. Rev. D*, **44**, 3124–3135.

Choptuik, M. W. (1993). Universality and scaling in gravitational collapse of a massless scalar field. *Phys. Rev. Lett.*, **2**, 9–12.

Chruściel, P. T. Corvino, J. and Isenberg, J. (2010). Construction of N-body initial data sets in general relativity. arXiv:1004.1355.

Clarke, C. J. S. (1998). Generalised hyperbolicity in singular space-times. *Class. Quant. Grav.*, **15**, 975–984.

Clarke, C. J. S. and Dray, T. (1987). Junction conditions for null hypersurfaces. *Class. Quantum Grav.*, **4**, 265–275.

Cordero, P. (1977). Canonical formulation of the spherically symmetric Einstein–Yang–Mills–Higgs system for a general gauge group. *Ann. Phys.*, **108**, 79–98.

Corichi, A. and Wilson-Ewing, E. (2010). Surface terms, asymptotics and thermodynamics of the Holst action. arXiv:1005.3298.

Crainic, M. and Fernandes, R. L. (2003). Integrability of Lie brackets. *Ann. Math.*, **157**, 575–620.

Damour, T. Henneaux, M. and Nicolai, H. (2003). Cosmological billiards. *Class. Quantum Grav.*, **20**, R145–R200.

Date, G. and Hossain, G. M. (2005). Genericity of big bounce in isotropic loop quantum cosmology. *Phys. Rev. Lett.*, **94**, 011302.

Date, G. Kaul, R. K. and Sengupta, S. (2009). Topological interpretation of Barbero–Immirzi parameter. *Phys. Rev. D*, **79**, 044008.

Deruelle, N. Sasaki, M. Sendouda, Y. and Yamauchi, D. (2009). Hamiltonian formulation of f(Riemann) theories of gravity. *Prog. Theor. Phys.*, **123**, 169–185.

Deser, S. Kay, J. H. and Stelle, K. S. (1977). Hamiltonian formulation of supergravity. *Phys. Rev. D*, **16**, 2448–2455.

DeWitt, B. S. (1967). Quantum theory of gravity. I. The canonical theory. *Phys. Rev.*, **160**, 1113–1148.

Dirac, P. A. M. (1958a). Generalized Hamiltonian dynamics. *Proc. Roy. Soc. A*, **246**, 326–332.

Dirac, P. A. M. (1958b). The theory of gravitation in Hamiltonian form. *Proc. Roy. Soc. A*, **246**, 333–343.

Dirac, P. A. M. (1969). *Lectures on Quantum Mechanics*. New York: Yeshiva University Press.

Dittrich, B. (2006). Partial and complete observables for canonical general relativity. *Class. Quant. Grav.*, **23**, 6155–6184.

Dittrich, B. (2007). Partial and complete observables for Hamiltonian constrained systems. *Gen. Rel. Grav.*, **39**, 1891–1927.

Dreyer, O. Krishnan, B. Shoemaker, D. and Schnetter, E. (2003). Introduction to isolated horizons in numerical relativity. *Phys. Rev. D*, **67**, 024018.

Eardley, D. M. (1998). Black hole boundary conditions and coordinate conditions. *Phys. Rev. D*, **57**, 2299–2304.

Eardley, D. M. and Smarr, L. (1979). Time functions in numerical relativity: Marginally bound dust collapse. *Phys. Rev. D*, **19**, 2239–2259.

Ellis, G. F. R. and Bruni, M. (1989). Covariant and gauge invariant approach to cosmological density fluctuations. *Phys. Rev. D*, **40**, 1804–1818.

Ellis, G. F. R. and MacCallum, M. A. H. (1969). A class of homogeneous cosmological models. *Commun. Math. Phys.*, **12**, 108–141.

Ellis, G. F. R. Murugan, J. and Weltman, A. (2011). *Foundations of Space and Time*. Cambridge University Press.

Fleischhack, C. (2009). Representations of the Weyl algebra in quantum geometry. *Commun. Math. Phys.*, **285**, 67–140.

Friedrich, H. (1996). Hyperbolic reductions for Einstein's equations. *Class. Quantum Grav.*, **13**, 1451–1469.

Friedrich, H. and Rendall, A. D. (2000). The Cauchy problem for the Einstein equations. *Lect. Notes Phys.*, **540**, 127–224.

Frittelli, S. (1997). Note on the propagation of the ocnstraints in standard $3+1$ general relativity. *Phys. Rev. D*, **55**, 5992–5995.

Frolov, V. P. and Vilkovisky, G. A. (1981). Spherically symmetric collapse in quantum gravity. *Phys. Lett. B*, **106**, 307–313.

Gambini, R. and Trias, A. (1980). Second quantization of the free electromagnetic field as quantum mechanics in the loop space. *Phys. Rev. D*, **22**, 1380–1384.

Garfinkle, D. (2002). Harmonic coordinate method for simulating generic singularities. *Phys. Rev. D*, **65**, 044029.

Garfinkle, D. (2004). Numerical simulations of generic singuarities. *Phys. Rev. Lett.*, **93**, 161101.

Garfinkle, D. and Gundlach, C. (1999). Symmetry-seeking spacetime coordinates. *Class. Quantum Grav.*, **16**, 4111–4123.

Geroch, R. and Winicour, J. (1981). Linkages in general relativity. *J. Math. Phys.*, **22**, 803–812.

Gibbons, G. W. and Hawking, S. W. (1977). Action integrals and partition functions in quantum gravity. *Phys. Rev. D*, **15**, 2752–2756.

Giesel, K. Hofmann, S. Thiemann, T. and Winkler, O. (2007). Manifestly gauge-invariant general relativistic perturbation theory: I. Foundations. arXiv:0711.0115.

Giles, R. (1981). The reconstruction of gauge potentials from Wilson loops. *Phys. Rev. D*, **24**, 2160–2168.

Girotto, F. and Saa, A. (2004). Semianalytical approach for the Vaidya metric in double-null coordinates. *Phys. Rev. D*, **70**, 084014.

Giulini, D. (1995). What is the geometry of superspace? *Phys. Rev. D*, **51**(10), 5630–5635.

Giulini, D. and Marolf, D. (1999). On the generality of refined algebraic quantization. *Class. Quantum Grav.*, **16**, 2479–2488.

Göckeler, M. and Schücker, T. (1989). *Differential Geometry, Gauge Theories, ad Gravity*. Cambridge University Press.

Goode, S. W. (1989). Analysis of spatially inhomogeneous perturbations of the FRW cosmologies. *Phys. Rev. D*, **39**, 2882–2892.

Govaerts, J. and Rashid, M. S. (1994). The Hamiltonian formulation of higher order dynamical systems. hep-th/9403009.

Gowdy, R. H. (1974). Vacuum spacetimes with two-parameter spacelike isometry groups and compact invariant hypersurfaces: Topologies and boundary conditions. *Ann. Phys.*, **83**, 203–241.

Grain, J. Cailleteau, T. Barrau, A. and Gorecki, A. (2010). Fully LQC-corrected propagation of gravitational waves during slow-roll inflation. *Phys. Rev. D*, **81**, 024040.

Grumiller, D. Kummer, W. and Vassilevich, D. V. (2002). Dilaton gravity in two dimensions. *Phys. Rept.*, **369**, 327–430.

Guth, A. H. and Pi, S. Y. (1985). Quantum mechanics of the scalar field in the new inflationary universe. *Phys. Rev. D*, **32**, 1899–1920.

Hartle, J. B. and Marolf, D. (1997). Comparing formulations of generalized quantum mechanics for reparametrization-invariant systems. *Phys. Rev. D*, **56**, 6247–6257.

Hawking, S. W. (1969). On the rotation of the universe. *MNRAS*, **142**, 129–141.

Hawking, S. W. (1971). Gravitational radiation from colliding black holes. *Phys. Rev. Lett.*, **36**, 1344–1346.

Hawking, S. W. (1972). Black holes in general relativity. *Commun. Math. Phys.*, **25**, 152–166.

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

Hawking, S. W. and Horowitz, G. T. (1996). The gravitational Hamiltonian, action, entropy and surface terms. *Class. Quantum Grav.*, **13**, 1487–1498.

Hawking, S. W. and Hunter, C. J. (1996). The gravitational Hamiltonian in the presence of non-orthogonal boundaries. *Class. Quantum Grav.*, **13**, 2735–2752.

Hayward, G. (1993). Gravitational action for spacetimes with nonsmooth boundaries. *Phys. Rev. D*, **47**, 3275–3280.

Hayward, S. A. (1994). General laws of black-hole dynamics. *Phys. Rev. D*, **49**, 6467–6474.

Hayward, S. A. (2006). Formation and evaporation of non-singular black holes. *Phys. Rev. Lett.*, **96**, 031103.

Heusler, M. (1998). Stationary black holes: Uniqueness and beyond. *Living Rev. Rel.*, **1**, 6.

Hiscock, W. A. Williams, L. G. and Eardley, D. M. (1982). Creation of particles by shell-focusing singularities. *Phys. Rev. D*, **26**, 751–760.

Hojman, S. A. Kuchař, K. and Teitelboim, C. (1976). Geometrodynamics regained. *Ann. Phys. (New York)*, **96**, 88–135.

Holst, S. (1996). Barbero's Hamiltonian derived from a generalized Hilbert–Palatini action. *Phys. Rev. D*, **53**, 5966–5969.

Horváth, Z. Kovács, Z. and Gergely, L. A. (2006). Geometrodynamics in a spherically symmetric, static crossflow of null dust. *Phys. Rev. D*, **74**, 084034.

Husain, V. (2009). Critical behaviour in quantum gravitational collapse. *Adv. Sci. Lett.*, **2**, 214.

Ikeda, N. (1994). Two-dimensional gravity and nonlinear gauge theory. *Ann. Phys.*, **235**, 435–464.

Ikeda, N. and Izawa, K. I. (1993). General form of dilaton gravity and nonlinear gauge theory. *Prog. Theor. Phys.*, **90**, 237.

Immirzi, G. (1997). Real and complex connections for canonical gravity. *Class. Quantum Grav.*, **14**, L177–L181.

Isham, C. J. (1983). Topological and global aspects of quantum theory. In: DeWitt, B. S. and Stora, R. (eds), *Relativity, Groups and Topology II*. Lectures given at the 1983 Les Houches Summer School on Relativity, Groups and Topology.

Israel, W. (1966). Singular hypersurfaces and thin shells in general relativity. *Nuovo Cimento B*, **44**, 1.

Iyer, V. and Wald, R. M. (1994). Some properties of the Noether charge and a proposal for dynamical black hole entropy. *Phys. Rev. D*, **50**, 846–864.

Jacobson, T. (2000). Trans-Planckian redshifts and the substance of the space-time river.

Jaramillo, J. L. and Gourgoulhon, E. (2010). Mass and angular momentum in general relativity. arXiv:1001.5429.

Kastrup, H. A. and Thiemann, T. (1994). Spherically symmetric gravity as a completely integrable system. *Nucl. Phys. B*, **425**, 665–686.

Keresztes, Z. and Gergely, L. Á. (2010). Covariant gravitational dynamics in $3 + 1 + 1$ dimensions. *Class. Quantum Grav.*, **27**, 105009.

Kiefer, C. (2004). *Quantum Gravity*. Oxford University Press.

Kiefer, C. Polarski, D. and Starobinsky, A. A. (1998). Quantum-to-classical transition for fluctuations in the early Universe. *Int. J. Mod. Phys. D*, **7**, 455–462.

Klauder, J. (2001). Quantization of constrained systems. *Lect. Notes Phys.*, **572**, 143–182.

Klauder, J. (2003). Affine quantum gravity. *Int. J. Mod. Phys. D*, **12**, 1769–1774.

Klauder, J. (2006). Overview of affine quantum gravity. *Int. J. Geom. Meth. Mod. Phys.*, **3**, 81–94.

Kobayashi, S. and Nomizu, K. (1963). *Foundations of Differential Geometry*. Vol. 1. New York: John Wiley.

Kobayashi, S. and Nomizu, K. (1969). *Foundations of Differential Geometry*. Vol. 2. New York: John Wiley.

Komar, A. (1959). Covariant conservation laws in general relativity. *Phys. Rev.*, **113**, 934–936.

Komar, A. (1979). Consistent factor ordering of general relativistic constraints. *Phys. Rev. D*, **20**, 830–833.

Kouletsis, I. Hájíček, P. and Bičák, J. (2003). Gauge-invariant Hamiltonian dynamics of cylindrical gravitational waves. *Phys. Rev. D*, **68**, 104013.

Kovács, Z. and Gergely, L. Á. (2006). New variables for brane-world gravity. *AIP Conf. Proc.*, **861**, 897–902.

Krishnan, B. and Schnetter, E. (2006). Non-symmetric trapped surfaces in the Schwarzschild and Vaidya spacetimes. *Phys. Rev. D*, **73**, 021502(R).

Kuchař, K. V. (1974). Geometrodynamics regained: A Lagrangian approach. *J. Math. Phys.*, **15**, 708–715.

Kuchař, K. V. (1976a). Dynamics of tensor fields in hyperspace. III. *J. Math. Phys.*, **17**, 801–820.

Kuchař, K. V. (1976b). Geometry of hypersurfaces. I. *J. Math. Phys.*, **17**, 777–791.

Kuchař, K. V. (1992). Time and interpretations of quantum gravity. In: Kunstatter, G. Vincent, D. E. and Williams, J. G. (eds), *Proceedings of the 4th Canadian Conference on General Relativity and Relativistic Astrophysics*. Singapore: World Scientific.

Kuchař, K. V. (1994). Geometrodynamics of Schwarzschild black holes. *Phys. Rev. D*, **50**(6), 3961–3981.

Langlois, D. (1994). Hamiltonian formalism and gauge invariance for linear perturbations in inflation. *Class. Quantum Grav.*, **11**, 389–407.

Langlois, D. and Vernizzi, F. (2005). Evolution of non-linear cosmological perturbations. *Phys. Rev. Lett.*, **95**, 091303.

Lau, S. R. (1996). New variables, the gravitational action and boosted quasilocal stress-energy-momentum. *Class. Quantum Grav.*, **13**, 1509–1540.

Lehner, L. (2001). Numerical relativity: A review. *Class. Quantum Grav.*, **18**, R25–R86.

Lewandowski, J. Newman, E. T. and Rovelli, C. (1993). Variations of the parallel propagator and holonomy operator and the gauss law constraint. *J. Math. Phys.*, **34**, 4646–4654.

Lewandowski, J. Okołów, A. Sahlmann, H. and Thiemann, T. (2006). Uniqueness of diffeomorphism invariant states on holonomy-flux algebras. *Commun. Math. Phys.*, **267**, 703–733.

Lichnerowicz, A. (1944). *J. Math. Pure Appl.*, **23**, 37.

Lidsey, J. E. Liddle, A. R. Kolb, E. W. Copeland, E. J. Barreiro, T. and Abney, M. (1997). Reconstructing the inflaton potential — an overview. *Rev. Mod. Phys.*, **69**, 373–410.

Lim, W. C. Andersson, L. Garfinkle, D. and Pretorius, F. (2009). Spikes in the Mixmaster regime of G_2 cosmologies. *Phys. Rev. D*, **79**, 123526.

Lovelock, D. (1970). The Einstein tensor and its generalizations. *J. Math. Phys.*, **12**, 498–501.

Lyth, D. H. (1985). Large scale energy density perturbations and inflation. *Phys. Rev. D*, **31**, 1792.

Maartens, R. (2004). Brane-world gravity. *Living Rev. Rel.*, **7**, 7. `http://www.livingreviews.org/lrr-2004-7`.

MacCallum, M. A. H. and Taub, A. H. (1972). Variational principles and spatially-homogeneous universes, including rotation. *Commun. Math. Phys.*, **25**, 173–189.

Madsen, M. S. (1988). Scalar fields in curved spacetimes. *Class. Quantum Grav.*, **5**, 627–639.

Malik, K. and Wands, D. (2004). Evolution of second-order cosmological perturbations. *Class. Quantum Grav.*, **21**, L65–L71.

Marolf, D. (1995). Refined algebraic quantization: Systems with a single constraint. gr-qc/9508015.

Mars, M. and Senovilla, J. (1993). Geometry of general hypersurfaces in spacetime: junction conditions. *Class. Quantum Grav.*, **10**, 1865–1897.

Martin, J. L. (1959). Generalized classical dynamics, and the 'classical analogue' of a Fermi oscillator. *Proc. Roy. Soc. A*, **251**, 536–542.

Mercuri, S. (2006). Fermions in Ashtekar-Barbero connections formalism for arbitrary values of the Immirzi parameter. *Phys. Rev. D*, **73**, 084016.

Misner, C. W. (1968). The isotropy of the Universe. *Astrophys. J.*, **151**, 431–457.

Misner, C. W. (1969). Mixmaster universe. *Phys. Rev. Lett.*, **22**, 1071–1074.

Modesto, L. (2006). Loop quantum black hole. *Class. Quantum Grav.*, **23**, 5587–5601.

Multamäki, T. Vainio, J. and Vilja, I. (2010). Hamiltonian perturbation theory in $f(R)$ gravity. *Phys. Rev. D*, **81**, 064025.

Nelson, W. and Sakellariadou, M. (2007a). Lattice refining loop quantum cosmology and inflation. *Phys. Rev. D*, **76**, 044015.

Nelson, W. and Sakellariadou, M. (2007b). Lattice refining LQC and the matter Hamiltonian. *Phys. Rev. D*, **76**, 104003.

Nielsen, A. B. (2009). Black holes and black hole thermodynamics without event horizons. *Gen. Rel. Grav.*, **41**, 1539–1584.

Nielsen, A. B. and Visser, M. (2006). Production and decay of evolving horizons. *Class. Quantum Grav.*, **23**, 4637–4658.
Novello, M. and Bergliaffa, S. E. P. (2008). Bouncing cosmologies. *Phys. Rep.*, **463**, 127–213.
Oppenheimer, J. R. and Snyder, H. (1939). On continued gravitational contraction. *Phys. Rev.*, **56**, 455–459.
Oriti, D. (2009). *Approaches to Quantum Gravity: Toward a New Understanding of Space, Time and Matter*. Cambridge University Press.
Penrose, R. (1963). Asymptotic properties of fields and space-times. *Phys. Rev. Lett.*, **10**, 66–68.
Penrose, R. (1965). Gravitational collapse and space-time singularities. *Phys. Rev. Lett.*, **14**, 57–59.
Penrose, R. (1990). Difficulties with inflationary cosmology. In: Fenves, E. (ed.), *Proceedings of the 14th Texas Symposium on Relativistic Astrophysics*. New York: N.Y. Acad. Sci.
Pfennig, M. J. and Ford, L. H. (1998). Scalar field quantum inequalities in static space-times. *Phys. Rev. D*, **57**, 3489–3502.
Pirani, F. A. E. and Schild, A. (1950). On the quantization of Einstein's gravitational field equations. *Phys. Rev.*, **79**, 986–991.
Pirani, F. A. E. Schild, A. and Skinner, R. (1953). Quantization of Einstein's gravitational field equations. II. *Phys. Rev.*, **87**, 452–454.
Podolský, J. and Svítek, O. (2005). Radiative spacetimes approaching the Vaidya metric. *Phys. Rev. D*, **71**, 124001.
Polarski, D. and Starobinsky, A. A. (1996). Semiclassicality and decoherence of cosmological perturbations. *Class. Quantum Grav.*, **13**, 377–392.
Pons, J. M. Salisbury, D. C. and Shepley, L. C. (1997). Gauge transformations in the Lagrangian and Hamiltonian formalisms of generally covariant theories. *Phys. Rev. D*, **55**, 658–668.
Pretorius, F. (2005). Evolution of binary black-hole spacetimes. *Phys. Rev. Lett.*, **95**, 121101.
Regge, T. and Teitelboim, C. (1974). Role of surface integrals in the Hamiltonian formulation of general relativity. *Ann. Phys.*, **88**, 286–318.
Rendall, A. D. (2005). *The Nature of Spacetime Singularities*. Singapore: World Scientific.
Rendall, A. D. and Weaver, M. (2001). Manufacture of Gowdy spacetimes with spikes. *Class. Quantum Grav.*, **18**, 2959–2975.
Roman, T. A. and Bergmann, P. G. (1983). Stellar collapse without singularities? *Phys. Rev. D*, **28**, 1265–1277.
Rovelli, C. (1991). What is observable in classical and quantum gravity? *Class. Quantum Grav.*, **8**, 297–316.
Rovelli, C. (2004). *Quantum Gravity*. Cambridge University Press.
Rovelli, C. and Smolin, L. (1990). Loop space representation of quantum general relativity. *Nucl. Phys. B*, **331**, 80–152.
Rovelli, C. and Smolin, L. (1995). Discreteness of area and volume in quantum gravity. *Nucl. Phys. B*, **442**, 593–619. Erratum: *Nucl. Phys. B* 456 (1995) 753.
Sachs, R. (1962). Asymptotic symmetries in gravitational theory. *Phys. Rev.*, **128**, 2851–2864.
Salopek, D. S. and Bond, J. R. (1990). Nonlinear evolution of long wavelength metric fluctuations in inflationary models. *Phys. Rev. D*, **42**, 3936–3962.

Saotome, R. Akhoury, R. and Garfinkle, D. (2010). Examining gravitational collapse with test scalar fields. arXiv:1004.3569.

Schaller, P. and Strobl, T. (1994). Poisson structure induced (topological) field theories. *Mod. Phys. Lett. A*, **9**, 3129–3136.

Schnetter, E. Krishnan, B. and Beyer, F. (2006). Introduction to dynamical horizons in numerical relativity. *Phys. Rev. D*, **74**, 024028.

Schwinger, J. (1962). Quantized gravitational field. *Phys. Rev.*, **130**, 1253–1258.

Senovilla, J. M. M. (1990). New class of inhomogeneous cosmological perfect-fluid solutions without Big-Bang Singularity. *Phys. Rev. Lett.*, **64**, 2219–2221.

Senovilla, J. M. M. (1998). Singularity theorems and their consequences. *Gen. Rel. Grav.*, **30**, 701–848.

Senovilla, J. M. M. (2007). A singularity theorem based on spatial averages. *Pramana*, **69**, 31–48.

Shibata, M. and Nakamura, T. (1995). Evolution of three-dimensional gravitational waves: Harmonic slicing case. *Phys. Rev. D*, **52**, 5428–5444.

Smarr, L. and York, J. W. (1978). Kinematical conditions in the construction of spacetime. *Phys. Rev. D*, **17**, 2529–2551.

Stewart, J. M. (1990). Perturbations of Friedmann–Robertson–Walker cosmological models. *Class. Quantum Grav.*, **7**, 1169–1180.

Strobl, T. (2004). Algebroid Yang–Mills theories. *Phys. Rev. Lett.*, **93**, 211601.

Thiemann, T. (1998a). Kinematical Hilbert spaces for fermionic and Higgs quantum field theories. *Class. Quantum Grav.*, **15**, 1487–1512.

Thiemann, T. (1998b). Quantum spin dynamics (QSD). *Class. Quantum Grav.*, **15**, 839–873.

Thiemann, T. (1998c). QSD V: Quantum gravity as the natural regulator of matter quantum field theories. *Class. Quantum Grav.*, **15**, 1281–1314.

Thiemann, T. (2007). *Introduction to Modern Canonical Quantum General Relativity*. Cambridge University Press.

Thiemann, T. and Kastrup, H. A. (1993). Canonical quantization of spherically symmetric gravity in Ashtekar's self-dual representation. *Nucl. Phys. B*, **399**, 211–258.

Thornburg, J. (1996). Finding apparent horizons in numerical relativity. *Phys. Rev. D*, **54**, 4899–4918.

Thornburg, J. (2007). Event and apparent horizon finders for $3 + 1$ numerical relativity. *Living Rev. Rel.*, **10**, 3.

Torre, C. G. (1999). Midi-superspace models of canonical quantum gravity. *Int. J. Theor. Phys.*, **38**, 1081–1102.

Uggla, C. van Elst, H. Wainwright, J. and Ellis, G. F. R. (2003). The past attractor in inhomogeneous cosmology. *Phys. Rev. D*, **68**, 103502.

Unruh, W. (1997). Time, gravity, and quantum mechanics. In: Savitt, S. F. (ed.) *Time's Arrows Today*. Cambridge University Press, 23–94.

Vachaspati, T. and Stojkovic, D. (2008). Quantum radiation from quantum gravitational collapse. *Phys. Lett. B*, **663**, 107–110.

Vaidya, P. C. (1953). 'Newtonian' Time in General Relativity. *Nature*, **171**, 260.

Vickers, J. A. and Wilson, J. P. (2001). Generalised hyperbolicity: Hypersurface singularities. gr-qc/0101018.

Vickers, J. A. and Wilson, J. P. (2000). Generalised hyperbolicity in conical space-times. *Class. Quantum Grav.*, **17**, 1333–1360.

Wald, R. M. (1984). *General Relativity*. University of Chicago Press.

Wald, R. M. and Iyer, V. (1991). Trapped surfaces in the Schwarzschild geometry and cosmic censorship. *Phys. Rev. D*, **44**, 3719–3722.

Wald, R. M. and Zoupas, A. (2000). General definition of “conserved quantities” in general relativity and other theories of gravity. *Phys. Rev. D*, **61**, 084027.

Wands, D. Malik, K. A. Lyth, D. H. and Liddle, A. R. (2000). A new approach to the evolution of cosmological perturbations on large scales. *Phys. Rev. D*, **62**, 043527.

Waugh, B. and Lake, K. (1986). Double-null coordinates for the Vaidya metric. *Phys. Rev. D*, **34**, 2978–2984.

Weiss, N. (1985). Constraints on Hamiltonian lattice formulations of field theories in an expanding universe. *Phys. Rev. D*, **32**, 3228–3232.

Winitzki, S. (2005). Drawing conformal diagrams for a fractal landscape. *Phys. Rev. D*, **71**, 123523.

York, J. W. (1971). Gravitational degrees of freedom and the initial-value problem. *Phys. Rev. Lett.*, **26**, 1656–1658.

York, J. W. (1986). Boundary terms in the action principles of general relativity. *Found. Phys.*, **16**, 249–257.

Index

BF-theory, 157
G-bundle, 280
n-bein, 282
2-dimensional gravity, 155

action principle, 22
adjoint action, 137
ADM momentum, 217
affine quantum gravity, 259
anchor map, 160, 286
angular momentum, 217
anisotropic shear, 128
  stress, 14, 167, 175
anisotropy parameter, 122
  plane, 125
  potential, 122
anomaly, 252, 253, 270
anti-commutation, 109
apparent horizon, 236
areal radius, 148
area increase, 240
Arnowitt–Deser–Misner (ADM) energy, 216
Ashtekar–Barbero connection, 93, 107, 108, 109, 256
Ashtekar variables, 98
associated vector bundle, 280
asymptotically anti de Sitter space-time, 206
  de Sitter space-time, 206
  empty space-time, 205
  flat space-time, 153, 205, 207, 211, 216
  Minkowski space-time, 206
  simple space-time, 206
asymptotic infinity, 203
asymptotic solution, 211
axial current, 106, 107

back-reaction, 244
background-field expansion, 38
background independence, 261
  metric, 255
Barbero–Immirzi parameter, 91, 106

Bardeen variables, 14, 169, 173
Belinskii–Khalatnokov–Lifschitz (BKL) conjecture, 126
Bianchi classification, 114, 119, 120, 129, 198
  class A, 119, 141, 145, 147
  class B, 119, 146, 147, 181
  identity, 18
  models, 131, 141, 277
  type I, 123, 128
  type IX, 124, 126, 128, 181
  type V, 181
black-hole entropy, 242
  evaporation, 244
  thermodynamics, 242
black hole, 82, 127, 198, 230, 235, 242, 271
Bondi–Metzner–Sachs (BMS) group, 214
boundary degree of freedom, 154
  term, 52, 53, 91, 146, 153, 181, 215
Brill–Lindquist initial data, 82
Brown–York quasilocal energy, 57, 216
BSSN equations, 78

canonical line element, 43, 165
Cartan's formula, 161
Carter–Penrose diagram, 204
Casimir–Darboux coordinates, 158, 283
Casimir function, 29, 31, 157, 158, 283
Cauchy horizon, 234
causal diagram, 201
caustic, 101, 191
centralizer, 137
chaos, 125, 127, 128
characteristic, 72
Christoffel symbols, 19, 39, 77, 83
Clifford algebra, 105
co-tetrad, 84
Codazzi equation, 47, 50
cohomology, 158
compatible presymplectic form, 158, 284, 285
complex connection, 108, 256

conformal completion, 204, 205, 208
diagram, 226, 228, 230, 233, 234, 244
gauge, 210
Hubble parameter, 10, 172
time, 9
transformation, 81, 175
conjugate point, 101, 192, 202
conjugation, 132, 137, 143
connection, 132, 134, 135, 256, 281
connection 1-form, 85, 86, 134, 135, 155, 280, 282
conservation equation, 176
of power, 177, 179, 271
consistency, 14, 15, 173, 176
consistent deformation, 254, 270
constraints, 18, 20, 22, 23, 167, 284
constraint algebra, 63, 170, 270
continuity equation, 9, 181
coordinate independence, 15
transformation, 14, 21, 66, 116, 166
cosmic censorship, 234
covariance, 52, 65
covariant derivative, 44, 85, 134, 276, 282
cross-normalization, 194, 211, 219
curvature, 87, 155
curvature 2-form, 87
curvature perturbation, 170, 177
potential, 124, 125, 181

Darboux coordinates, 27
deceleration, 10, 11
parameter, 9
densitized triad, 93, 171, 172
density weight, 37, 93, 141, 151, 276
deparameterization, 70, 103, 123, 260, 265
derivative expansion, 176
determinant, 39, 43
DeWitt supermetric, 113
de Sitter space, 9, 10
diagonal Bianchi model, 121
diffeomorphism, 63, 64
constraint, 51, 61, 81, 95, 96, 99, 102, 104, 110, 148, 151, 159, 171, 173, 176, 277
dilaton, 154
gravity, 154, 156
potential, 155
Dirac action, 106
bracket, 31, 97, 285
classification, 285
equation, 106
observable, 34, 71, 152
quantization, 251
discrete spectrum, 258, 262, 266
distortion, 80
driver, 80
dust, 10, 101
Hamiltonian, 103
dyad, 155, 282

Eddington–Finkelstein coordinates, 208, 223
effective constraint, 253, 271
equation, 263
Hamiltonian, 265, 267
Einstein–Hilbert action, 5, 38, 50, 54
Einstein static universe, 15, 204
tensor, 20, 155, 164
electric field, 103
electromagnetism, 37, 103
energy condition, 10, 181, 237, 241, 268, 271
current, 99, 103, 175
density, 8, 98, 101, 104, 175
epsilon tensor density, 276
equation of state, 10, 104
Euclidean group, 132, 250
Euler–Lagrange equations, 22
Eulerian observer, 42, 92, 98, 99, 113
Euler angles, 182
event horizon, 208, 231, 234, 236, 238, 271
evolving observable, 71
expansion, 11, 180, 185, 192
exponential map, 114, 278
extended phase space, 67, 170
extrinsic-curvature tensor, 33, 45, 50, 53, 57, 93, 94, 108, 171, 180, 185, 191, 193
extrinsic geometry, 45

factor ordering, 253
space, 132, 194
fermion, 83, 104, 106, 107, 109
fiber bundle, 279
field strength, 103, 140
field theory, 35, 255, 277
first-class Hamiltonian, 34
constraint, 20, 30, 34, 51, 61, 66, 97, 173, 249, 252, 259, 270, 285
constraint surface, 30
formulation, 82, 106, 155
first-class constraint, 159
first structure equation, 87, 90, 156
flux, 257
focal point, 191, 192, 195
foliation, 29, 40, 64, 92, 98
frame, 8, 66, 69, 130, 149, 173
frame independence, 174
free group action, 277
Friedmann–Lemaître–Robertson–Walker (FLRW) models, 4, 42, 88, 123, 142, 144, 166, 209, 222
Friedmann equation, 7, 9, 128, 176
functionally differentiable, 154

functional derivative, 36, 154
differentiability, 36, 52
future null infinity, 205
timelike infinity, 204

gamma matrices, 105
gauge, 135
gauge-invariant perturbation, 168, 172
variable, 174
vector mode, 173
gauge-source function, 77
gauge conditions, 166
fixing, 38, 69, 76, 148, 149
freedom, 6
invariance, 252
transformation, 31, 34, 51, 66, 133, 136, 153, 157, 161, 168, 172, 280, 282, 286
Gauss constraint, 94, 96, 104, 108, 110, 159
equation, 47, 50
law, 37
generalized harmonic coordinates, 77
general covariance, 20
generator of gauge transformations, 68
geodesic congruence, 11, 185
deviation equation, 191
Gibbons–Hawking term, 54, 154
global section, 135
Gowdy models, 127
Grassmann variable, 109
gravitational collapse, 226, 227, 234, 238
gravitational waves, 168
ground state, 179
groupoid, 279
group averaging, 251

half-densitized spinor, 109
Hamiltonian, 24
Hamiltonian constraint, 51, 61, 81, 95, 96, 98, 102, 104, 121, 122, 151, 171, 176
equations of motion, 6, 7, 26
vector field, 27, 28, 283
harmonic coordinates, 77
Hawking radiation, 244, 271
higher-curvature action, 61, 254, 270
higher-derivative Lagrangian, 60
higher-dimensional theory, 61
hole argument, 21
holonomy, 256, 281
correction, 266
Holst–Dirac action, 107
Holst action, 91, 106
homogeneity, 4, 118, 129, 130, 131, 141, 142, 146, 198
homotopic, 162
horizon, 226, 234, 235
horizontal space, 281
Hubble exit, 271
parameter, 4, 9, 175, 180
radius, 174, 178
hyperbolic, 73, 77, 166
reduction, 77
hypersurface-deformation algebra, 65, 163, 254, 270, 286
hypersurface tilt, 65

induced metric, 41
inflation, 128, 174, 178
inhomogeneity, 126, 174
inhomogeneous model, 130
initial-value problem, 10, 18, 20, 72, 234
initial data, 234
internal time, 70, 71, 265
vector, 84
intrinsic-curvature tensor, 44
invariance up to gauge, 136
invariant 1-forms, 118
connection, 131, 136, 139, 142, 144, 182
densitized triad, 145
metric, 129, 130, 141
tensor, 118, 130, 141
vector field, 115, 116
inverse-triad correction, 267
isolated horizon, 237
isotropic connection, 143
isotropization, 128
isotropy, 4, 132
isotropy subgroup, 132, 137, 138, 142

Jacobi identity, 27, 115, 119, 157, 180, 283
vector field, 191
junction conditions, 221, 224

Kantowski–Sachs model, 129, 198
Kasner epoch, 125
exponent, 124
metric, 124
solution, 124
kinematical Hilbert space, 251, 255
phase space, 249
Kirillov–Kostant structure, 157, 284
Klein–Gordon equation, 8, 13, 16, 73, 178, 260

lapse function, 4, 42, 51, 67, 76, 170
large-scale mode, 176
lattice refinement, 268
leaf-symplectic structure, 158
left-invariant vector field, 277
left-translation, 117, 277
Legendre transformation, 24, 60, 102, 277
Leibniz rule, 27, 85, 160

Lichnerowicz equation, 82
Lie algebra, 115, 131, 277
algebroid, 64, 160, 163, 286
algebroid morphism, 161, 287
derivative, 63, 64, 93, 103, 116, 166, 169, 274, 276, 277
group, 114, 277
groupoid, 163, 287
linearized equations, 166
perturbations, 163
linear perturbation, 12, 14
perturbation equation, 173
locally rotationally symmetric (LRS) models, 142
trivial, 133, 279
local gauge transformation, 134
section, 135
trivialization, 135
loop quantum cosmology, 261, 266
gravity, 256
Lorentz connection, 106, 107
transformation, 86, 105

magnetic charge, 138
field, 104, 140
monopole, 140
marginal trapped surface, 236
mass, 152
matching, 217, 227, 241, 271
matrix exponential, 278
group, 277
Maurer–Cartan form, 131, 135, 139, 141, 142, 182, 278
relations, 118, 278
maximal slicing, 80, 111
torus, 138, 143
maximum arc-length, 186, 190
metric perturbation, 165
midi-superspace model, 129, 144
mini-superspace, 129
minimal-distortion shift, 80
Minkowski space, 37, 56, 65, 84, 203, 219
Misner variables, 122
mixmaster behavior, 125
mode decomposition, 165
moments, 263
Mukhanov–Sasaki equation, 178
musical isomorphism, 28

naked singularity, 234
non-linear inhomogeneity, 175
non-minimal coupling, 7, 100, 106, 109
normal, 41
null-junction conditions, 222, 227
null coordinates, 203
dust, 101, 227, 231
infinity, 230
normal, 219
singularity, 233
surfaces, 219
vector, 194
numerical relativity, 76, 77, 78, 237

observable, 152, 249, 252
off-shell, 61
on-shell, 64
Oppenheimer–Snyder model, 222, 231
orbit, 129, 132
Ostrogradsky procedure, 60

Painlevé–Gullstrand metric, 182, 200
Palatini action, 93, 106
formulation, 82
parallelizable, 277
parallel transport, 85, 281
parity, 105
past null infinity, 205
timelike infinity, 204
perfect fluid, 9, 10, 100, 102
perturbed line element, 169
physical Hilbert space, 259
inner product, 251, 260
state, 251, 259
Planck length, 7, 244, 258, 268
Poincaré group, 214
transformation, 66
Poisson bracket, 26, 36, 63, 108, 109, 121, 263, 283
equation, 14
manifold, 156, 157, 265
sigma model, 157, 287
structure, 28, 159
tensor, 27, 109, 156, 160, 283
positive/negative frequency, 260
pressure, 8, 99
presymplectic form, 29, 284
manifold, 285
primary constraint, 6, 24, 50, 60, 159
constraints, 67
primordial spectrum, 179
principal fiber bundle, 135, 280
part, 166
symbol, 72
problem of time, 70
proper time, 4, 123
pseudoscalar, 106
pull back, 30, 275
bundle, 287
push-forward, 116, 275

quantization, 108, 178, 179, 250, 255, 259
quantum back-reaction, 268
  correction, 264
  cosmology, 258
  field theory, 109, 243
  fluctuation, 174, 178, 179, 263
  geometry, 258
  gravity, 245
  phase space, 264
quasi-linear, 72
quasilocal, 56

radiation, 10
Raychaudhuri equation, 8, 9, 176, 185, 193, 237, 268
reality condition, 98, 108, 267
reduced bundle, 137
  Hamiltonian, 154
  manifold, 130
  phase space, 31, 153, 249
  structure group, 139, 143, 144
  symplectic structure, 29
relational, 70
relativistic particle, 71, 260
Ricci equation, 48
  rotation 1-forms, 88
    coefficients, 155
  scalar, 5, 118
  tensor, 19, 39, 155, 164
Riemann tensor, 87, 89, 90, 91, 155
right-translation, 117, 277
right action, 135, 280
Robinson–Trautman metric, 235
rotation, 185

scalar density, 37, 276
  field, 7, 13, 62, 99, 255
  mode, 14, 165, 168, 169
scale-free, 180
scale factor, 4, 144, 181
Schrödinger equation, 260, 264
Schwarzschild space-time, 57, 150, 198, 208, 223, 271
second-class constraint, 30, 31, 32, 38, 61, 69, 96, 97, 107, 148, 285
    surface, 30
second-order formulation, 97, 106
secondary constraint, 6, 32, 51, 67, 94, 96, 159, 171
second structure equation, 87, 155
section, 133, 279
semiclassical state, 253, 268
semidirect product, 132, 142
separate-universe approach, 175, 186
shear, 80, 128, 180, 185
shift vector, 42, 51, 67, 76, 170
sigma model, 157
singularity, 9, 10, 11, 79, 123, 124, 126, 186, 198, 226, 230, 233, 244, 268, 271
  theorem, 12, 181, 193, 195, 197, 202, 235, 268
slicing conditions, 79
smearing, 36, 61, 255
space-time gauge, 67, 76
  manifold, 76
spacelike infinity, 205
spatial tensor, 44
  infinity, 215
  stress tensor, 99
spherically symmetric connection, 136
  electromagnetism, 140
spherical symmetry, 130, 141, 144, 147, 151, 155
  trapped surface, 199, 200, 231
spikes, 127
spinor, 104
spin connection, 93, 107, 283
sructure function, 64
state, 179
stationary horizon, 240
stiff fluid, 100, 129
stress-energy, 104
  tensor, 8, 98, 99, 101, 177
strictly hyperbolic, 73
structure constants, 115, 117, 119, 141, 159, 180, 277
  formation, 269
  functions, 160, 170, 286
  group, 131, 134, 280
super-translation, 214
supergravity, 61, 109
surface delta function, 219
  gravity, 241
  step function, 219
symmetric criticality, 6, 146
  hyperbolic, 73
  principal fiber bundle, 137, 141
symmetry group, 114, 129, 131, 138
  reduction, 129
symplectic form, 28, 283
  leaf, 30, 157, 283
  structure, 28, 284

target space, 157
tensor density, 37, 118, 276
  mode, 165, 168, 174
tetrad, 84, 106, 282
  formulation, 90
time, 35, 40, 51, 66, 70
time-evolution vector field, 42, 66, 69, 100
time derivative, 43
  gauge, 92

topological charge, 138, 140, 144, 145
torsion, 83, 87, 107, 156
total Hamiltonian, 6, 26, 32, 34, 51, 56, 69
transition functions, 279
transitive group action, 114, 129, 131, 135, 277
transversal Lie algebra, 286
transverse-traceless tensor, 165
transverse vector, 165
trapped region, 226
  surface, 195, 198, 236
trapping horizon, 236, 238, 271
triad, 92, 93, 282
  formulation, 97
trivial bundle, 133

uncertainty relation, 243

Vaidya solution, 227
variation, 36, 52, 60, 82, 90, 145
vector bundle, 84, 133, 279
  current, 105
  mode, 165, 168
  potential, 103
vertical vector field, 280

wave equation, 73, 76, 78, 167, 168
weak equality, 26
Weyl group, 138, 143
Wheeler–DeWitt quantization, 256
world-sheet, 157

Yang–Mills monopole, 145
  theory, 157