Daniel Siemssen

The Semiclassical Einstein Equation on Cosmological Spacetimes

Ph.D. Thesis

Dipartimento di Matematica
Università degli Studi di Genova

The Semiclassical Einstein Equation on Cosmological Spacetimes

Ph.D. thesis submitted by Daniel Siemssen
Genova, February 2015

Dipartimento di Matematica
Università degli Studi di Genova

Supervisor: Prof. Dr. Nicola Pinamonti
Examiner: Prof. Dr. Valter Moretti

Abstract

The subject of this thesis is the coupling of quantum fields to a classical gravitational background in a semiclassical fashion. It contains a thorough introduction into quantum field theory on curved spacetime with a focus on the stress-energy tensor and the semiclassical Einstein equation. Basic notions of differential geometry, topology, functional and microlocal analysis, causality and general relativity will be summarised, and the algebraic approach to quantum field theory on curved spacetime will be reviewed. The latter part contains an introduction to the framework of locally covariant quantum field theory and relevant quantum states: Hadamard states and, on cosmological spacetimes, adiabatic states. Apart from these foundations, the original research of the author and his collaborators will be presented:

Together with Fewster, the author studied the up and down structure of circular and linear permutations using their decomposition into so-called atomic permutations. The relevance of these results to this thesis is their application in the calculation of the moments of quadratic quantum fields in the quest to determine their probability distribution.

In a work with Pinamonti, the author showed the local and global existence of solutions to the semiclassical Einstein equation in flat cosmological spacetimes coupled to a massive conformally coupled scalar field by solving simultaneously for the quantum state and the Hubble function in an integral-functional equation. The theorem is proved with the Banach fixed-point theorem using the continuous functional differentiability and boundedness of the integral kernel of the integral-functional equation.

Since the semiclassical Einstein equation neglects the quantum nature of the stress-energy tensor by ignoring its fluctuations, the author proposed in another work with Pinamonti an extension of the semiclassical Einstein equations which couples the moments of a stochastic Einstein tensor to the moments of the quantum stress-energy tensor. In a toy model of a Newtonianly perturbed exponentially expanding spacetime it is shown that the quantum fluctuations of the stress-energy tensor induce an almost-scale-invariant power spectrum for the perturbation potential and that non-Gaussianties arise naturally.

Contents

Introduction

The subject of this thesis is the interplay between quantum matter and gravity, *i.e.*, the coupling of quantum fields to a classical gravitational background in a *semiclassical* fashion. Semiclassical gravity describes physics midway between the classical regime covered by the Einstein equation and a full-fledged quantum gravity. However, while no theory of quantum gravity is universally accepted, quantum field theory on curved spacetimes offers an approach to the 'low'-energy and 'small'-curvature regime based on the firm foundation of quantum field theory and Lorentzian geometry. Despite of the employed approximation, it has made several successful and relevant predictions like the Fulling–Davies–Unruh effect [67, 105, 214], the Hawking effect [100, 118], cosmological particle creation [171] and the generation of curvature fluctuations during inflation [108, 120, 161, 202]. It is generally expected that a successful theory of quantum gravity also describes these phenomena and, in fact, they are used as criteria to select candidate theories.

Although quantum gravity is one motivation for studying quantum field theory on curved spacetime, it is not the only reason. While quantum field theory is typically formulated on a Minkowski background, the Universe appears well-described by a curved spacetime and Minkowski spacetime provides only a local approximation. However, even the slightest gravitational interaction causes many of the basic assumptions of 'standard' quantum field theory on Minkowski spacetime to fail and in important situations, like inflation, the departure from a flat background is not small but causes important effects that cannot be neglected. From this point of view it would be conceptually very unsatisfying if it was not possible to successfully formulate quantum field theory on a curved spacetime in such a way that it reduces to standard QFT in the case of a flat background.

Attempts to formalize quantum field theory in a mathematically exact manner have led to many significant insights into the structure of quantum fields: the CPT theorem, the spin-statistics connection, and superselection sectors to name a few, see *e.g.* [110, 206]. By studying aspects of semiclassical gravity and quantum field theory on curved spacetimes in the rigorous framework of algebraic quantum field theory, one hopes to gain deep and novel insights into the subtle nature of quantum fields on curved spacetimes and at the same time often prove theorems that have also a purely mathematical value. Moreover, as a consequence of the correspondence principle, it is highly plausible that a careful investigation of the semiclassical theory gives us further hints about the structure of an eventual theory of quantum gravity. In particular, one can expect that observations in cosmology are already described to high precision within semiclassical Einstein gravity and that tight limits can be placed on the creation of extreme objects such as wormholes in generic spacetimes.

In the formulation of quantum field theory on Minkowski spacetime one usually starts the with the unique Poincaré-invariant vacuum state as the ground state in a Fock space motivated by the particle interpretation. On a generic spacetime, due to the absence of any symmetries, no such distinguished state can exist and, as illustrated in the Unruh and the Hawking effect, no unique particle interpretation is

available. This suggests that the starting point for a quantum field theory on curved spacetimes should be a formulation that does not require a preferred state. For this reason, rigorous quantum field theory on curved spacetimes is often discussed within the *algebraic approach* to quantum field theory [109–111]. In the algebraic approach one begins by considering an abstract algebra of quantum fields, which respects conditions of locality (quantum fields only depend on the local structure of the spacetime) and causality (causally separated quantum fields (anti)commute).

A modern formulation of QFT on curved spacetimes is the framework of *locally covariant quantum field theory* [46]. It helped the development of several important contributions to QFT on curved spacetimes like renormalization and perturbative algebraic quantum field theory [43, 101, 102], superselection sectors on curved spacetimes [47], abstract and concrete results on gauge theories [30–32, 59, 86, 194] and many other results. In this framework one considers quantum field theories as covariant functors from a category of background structures to a category of physical systems. In most simple examples the background structure is given by a category of globally hyperbolic spacetimes with so-called hyperbolic embeddings as morphisms, but the background structure can be replaced by anything reasonable that allows for a categorical formulation, see [2, 177] for examples of alternative choices. A suitable category representing physical systems for algebraic quantum field theory is a category of $*$-algebras so that a quantum field theory maps a spacetime to an algebra of observables in that spacetime.

However, while states, *viz.*, positive linear functionals on a $*$-algebra, are not necessary for the formulation of the theory, they are indispensable if one wants to make quantitative predictions. Given a state it is possible to return to a Hilbert space picture as a representation of the $*$-algebra of observables via the Gel'fand–Naimark–Segal theorem. Not all possible states on the algebra of quantum fields are of equal physical importance. Physically and mathematically preferred states are the so-called Hadamard states, which have an ultraviolet behaviour analogous to that of the Minkowski vacuum. Hadamard states are for example required for a reasonable semiclassical Einstein equation; otherwise the fluctuations of the quantum stress-energy tensor are not even distributions and the semiclassical Einstein equation becomes physically meaningless, because we equate a quantity with a probabilistic interpretation and 'diverging' fluctuations with a classical non-fluctuating quantity. Major advances in quantum field theory on curved spacetime were achieved after it was realized in [181] that all Hadamard states satisfy a constraint on the wavefront set of the n-point functions of the state. This constraint was called microlocal spectrum condition in allusion to the condition from Wightman quantum field theory on Minkowski spacetime. In particular, this discovery led to the formulation of a rigorous theory of renormalization and a concept of normal ordering on curved spacetimes [44, 45, 123, 124].

The developments of quantum field theory on curved spacetimes were often driven by problems related to the *semiclassical Einstein equation*. In the semiclassical Einstein equation contains instead of a classical stress-energy tensor the expectation value of a quantum stress-energy tensor $:T_{ab}:$ in a certain state ω:

$$G_{ab} + \Lambda g_{ab} = \frac{8\pi G}{c^4} \omega(:T_{ab}:).$$

The quantum stress-energy tensor may be obtained by replacing the products of classical fields in the classical stress-energy tensor by normally ordered products of quantum fields. This requires the notion of normal ordering on curved spacetimes mentioned above. The resulting quantum stress-energy is not uniquely fixed but, due to the non-uniqueness of the normal ordering prescription, subject to a renormalization freedom, which is a polynomial of local geometric quantities, whose coefficients are called renormalization constants.

Quantum field theory on curved spacetimes is best understood in a few special cases of highly symmetric spacetimes. In particular quantum fields on Friedmann–Lemaître–Robertson–Walker spacetimes are well-studied as they are important in quantum cosmology. Nevertheless, already in this simplified case many interesting effects occur, for example the creation of particles in an expanding spacetime [171]. Due to the developments discussed above, in recent years computations of quantum field theoretic effects in cosmological spacetimes and their backreaction to the spacetime via the semiclassical Einstein equation have come into the reach of the algebraic approach to QFT on curved spacetimes.

A first step towards doing cosmology in algebraic QFT on curved spacetimes is often the construction of appropriate states. Noteworthy recent works are the holographic (or bulk-to-boundary) construction [57, 61–65, 1, 157, 159] and the states of low energy [165, 210] (see also [70, 71]) which are a Bogoliubov transformation of adiabatic states [134, 146, 172]. Given a state, one can study the semiclassical Einstein equation to study the backreaction effects of quantum matter fields; this has been done, for example, in [56, 66, 112, 114]. Going one step further, one can attempt to solve this semiclassical Einstein equation, *i.e.*, finding a spacetime and a state on that spacetime so that the equation holds. This problem was analyzed for cosmological spacetimes in [83, 178, 3]. Other works studied linearized gravity [92, 113], inflation [81] (see also [4] for a non-standard approach) and other cosmological models [221] in the algebraic framework. Furthermore, several researches have studied thermal aspects of quantum fields on curved spacetime, [49, 82, 196, 197] to name a few, which are arguably of importance to quantum cosmology.

In this thesis several aspects of the works cited above will be summarized and, when necessary, developed further. To give this work a clearer structure, it is divided into three parts.

The first part is mostly intended to lay the foundations of the remaining two parts. In Chap. 1 a rapid summary of subjects from differential geometry relevant to QFT on curved spacetimes is presented but it also contains a few sections and remarks on subjects which are usually not covered in standard text books on differential geometry, *e.g.* bitensors. Chap. 2 focuses on the particular case of Lorentzian geometry including notions of causality, the classical Einstein equation and cosmology. Analysis, in the broadest sense, will be the subject of Chap. 3 and in that chapter various results on topology, $*$-algebras, functional derivatives and their relation to the Banach fixed-point theorem, microlocal analysis and wave equations will be summarised. In favour of not jumping back and forth between different subjects in these three sections I chose a rather unpedagogical order and the reader should be aware that there are many interrelations between the various sections. This should, however, not be a too large an obstacle for the reader. Chap. 4 concerns the enumerative combinatorics

of permutations and appears somewhat unrelated to most of this thesis. However, combinatorics is very important in many applications of quantum field theory and the results presented in this chapter are important in the moment problem for quadratic quantum fields [90]. The contents of this last chapter represent work by the author in collaboration with Fewster and were published in [95].

In the second part of this thesis several aspects of quantum field theory on curved spacetimes will be discussed. It begins with an introduction to the categorical framework of locally covariant quantum field theory with an emphasize on the field algebra of Klein–Gordon-like quantum fields in Chap. 5. In Chap. 6 we discuss quantum states and in particular the construction of adiabatic states of cosmological spacetimes and the holographic construction of Hadamard states on asymptotically flat spacetimes.

The third and last part of this thesis represents the largest portion of novel research done during the authors Ph.D. studies; here the semiclassical Einstein equation will be analyzed in detail. The basic notions including the stress-energy tensor for the scalar field and its renormalization will be introduced in Chap. 7. In Chap. 8 the proof of the author and Pinamonti [3] on the local and global existence of solutions to the semiclassical Einstein equation will be presented. Finally, in the last chapter of this thesis (Chap. 9), the fluctuations of the stress-energy tensor will be analyzed and how their backreaction to the metric may be accounted for; this work in collaboration with Pinamonti was already presented in [4].

This thesis will be concluded with some final remarks on the research presented above in the conclusions. Following this the cited references may be found in the bibliography, starting with the works (co)authored by the present author.

I

Foundations

Zudem ist es ein Irrtum zu glauben, daß die Strenge in der Beweisführung die Feindin der Einfachheit wäre. An zahlreichen Beispielen finden wir im Gegenteil bestätigt, daß die strenge Methode auch zugleich die einfachere und leichter faßliche ist. Das Streben nach Strenge zwingt uns eben zur Auffindung einfacherer Schlußweisen; auch bahnt es uns häufig den Weg zu Methoden, die entwickelungsfähiger sind als die alten Methoden von geringerer Strenge.

— David Hilbert, "Mathematische Probleme" (1900), p. 257.

<div style="text-align: right; font-size: 3em;">1</div>

Differential geometry

Summary

This chapter is *mostly* a summary of some common definitions and standard results on differential geometry and most of its content can be safely skipped by a reader well acquainted with the topic. Proofs are omitted everywhere except in the last section and may be found in any text book on differential geometry. Nevertheless, the author has attempted to present the material in such a way that many statements should become self-evident, although, as always, care should be taken.

In the first section (Sect. 1.1) the basic theory of differentiable manifolds and vector bundles is summarized. Here the notions of coordinates, maps between manifolds, vector bundles and sections, the (co)tangent bundle, (co)vectors and (co)vector fields, curves, tensor and exterior tensor product bundles, bundle metrics, frames, differential operators, and the index notation are explained. The second section (Sect. 1.2) is concerned with the definition of connections on vector bundles and the objects that follow from this. That is, it discusses the notions of curvature, geodesics, and the slightly unrelated concept of Killing vector fields. Differential forms and integration are introduced in the third section (Sect. 1.3). In particular we will introduce the de Rham cohomology, the Hodge star and the dual of the exterior derivative, the codifferential, which leads to the definition of the Laplace–de Rham operator, and close with a short discussion of integral manifolds.

In the presentation of these three sections the author follows partially that of [142] and also [6], but these standard definitions may be found in many places in the literature.

The fourth and last section (Sect. 1.4) treats a more obscure topic: bitensors. Bitensors are already introduced in an abstract manner in Sect. 1.1.6; a simple, yet important, example are biscalars: functions on the product $M \times M$ of a manifold M. In this last section concrete and important cases of bitensors such as Synge's world function and the van Vleck–Morette determinant are discussed. It will also form the foundation for the discussion of the Hadamard coefficients in Sect. 6.1.3. An excellent resource on bitensors is the review article [179], which contains most of the first part of this section. The second part of this section is concerned with computational methods that help the calculation of coincidence limits of bitensors. Here we will discuss the semi-recursive Avramidi method developed in [169]. We close this section with a recursive method to calculate the coefficients of an asymptotic expansion of Synge's world function in coordinate separation. To the authors knowledge, this simple and efficient method has never been fully developed but traces of it may be found in [168].
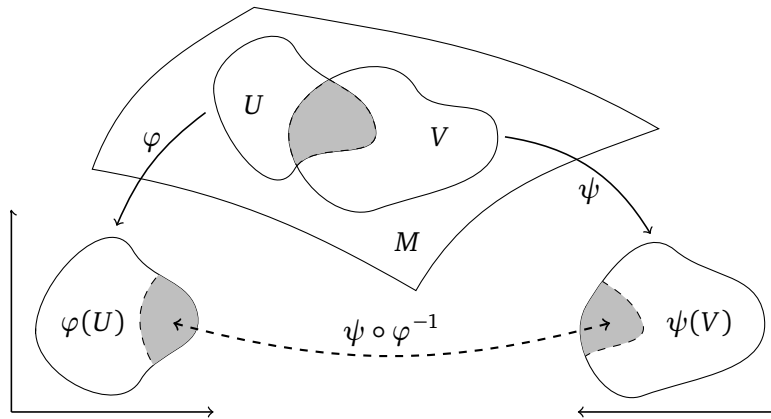
Figure 1.1. Two overlapping charts and their transition map.

1.1   Differentiable manifolds and vector bundles

A *topological manifold of dimension n* is a second-countable Hausdorff space M that is locally homeomorphic to $\mathbb{R}^n$ (*i.e.*, each point of M has a neighbourhood that is homeomorphic to an open subset of $\mathbb{R}^n$). We often omit the dimension of the manifold and simply say: M is a *topological manifold*.

Since a topological manifold M is locally homeomorphic to $\mathbb{R}^n$, we can assign coordinates to points of M in each open neighbourhood $U \subset M$: A *(coordinate) chart* of M is a pair (U, φ) which gives exactly such a homeomorphism

$$\varphi : U \to \varphi(U) \subset \mathbb{R}^n;$$

we call U its *coordinate neighbourhood*. The component functions $(x^1, \dots, x^n) = \varphi$ are called *(local) coordinates* on U.

An *atlas A* of M is a family of charts $(U_i, \varphi_i)_{i \in \mathbb{N}}$ which cover M. If any two overlapping charts $(U, \varphi), (V, \psi)$ in an atlas are *smoothly compatible*, *viz.*, the *transition map*

$$\psi \circ \varphi^{-1} : \varphi(U \cap V) \to \psi(U \cap V)$$

is a smooth, bijective map with a smooth inverse (Fig. 1.1), we say that the atlas is *smooth*. We further say that a smooth atlas A on M is *maximal* if it is not properly contained in any larger smooth atlas so that any chart which is smoothly compatible with the charts of A is already contained in A.

Finally, a *smooth manifold* is a pair (M, A), where M is a topological manifold and A a maximal smooth atlas. A maximal smooth atlas might not exist and, if it exists, it is not necessarily unique as shown, *e.g.*, by the existence of exotic $\mathbb{R}^4$. Nevertheless, usually a canonical smooth atlas is understood from context. Then we omit the explicit mention of the maximal smooth atlas A and say: M is a smooth manifold. One can replace the requirement of the transition maps in an atlas to be smooth by requiring that the transition maps are C^k, (real-)analytic or complex-analytic (if $\dim M = 2n$, we have $\mathbb{R}^{2n} \simeq \mathbb{C}^n$) thus arriving at the notions of C^k, (real-)analytic and complex-analytic manifolds.
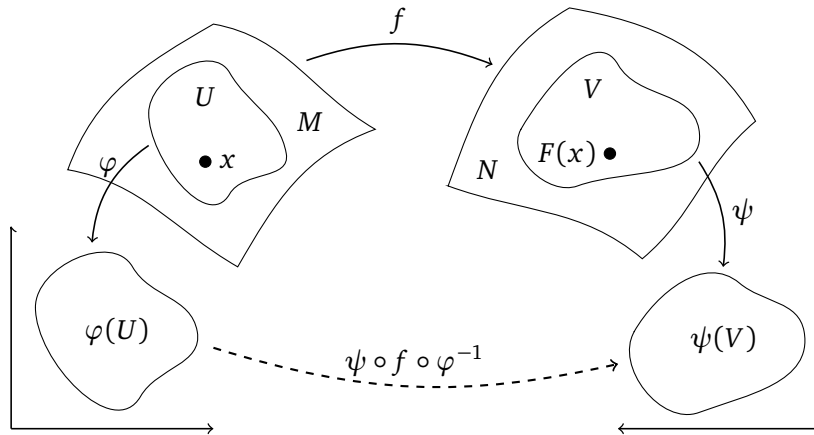
Figure 1.2. A map between two manifolds.

1.1.1 Smooth maps

A real-valued function $f : M \to \mathbb{R}$ on a smooth manifold M is a C^k, *smooth or analytic function* if there exists a chart (U, φ) containing x at every $x \in M$ such that the composition $f \circ \varphi^{-1}$ is C^k, smooth or analytic on the image $\varphi(U)$; the spaces of these function are denoted $C^k(M)$, $C^\infty(M)$ and $C^\omega(M)$ respectively.

More generally, a map $f : M \to N$ between two smooth manifolds M and N is C^k, smooth or analytic, if there exist charts (U, φ) at $x \in M$ and (V, ψ) at $f(x) \in N$ such that the composition

$$\psi \circ f \circ \varphi^{-1} : \varphi(U) \to \psi(V)$$

is C^k, smooth or analytic (Fig. 1.2). If M and N have equal dimension and f is a homeomorphism such that f and its inverse f^{-1} are smooth, we call $f : M \to N$ a *diffeomorphism*. Whenever there exists such a diffeomorphism between, they are *diffeomorphic*; in symbols $M \simeq N$.

1.1.2 Vector bundles

A (smooth) $\mathbb{K}$-*vector*[1] *bundle* of dimension n

$$\pi : E \to M$$

consists of two smooth manifolds E, the *total space*, and M, the *base (space)*, and a smooth surjection π, the *bundle projection*, that associates to every $x \in M$ a n-dimensional $\mathbb{K}$-vector space $E_x = \pi^{-1}(\{x\})$, the *fibre* of E at x (Fig. 1.3). Moreover, we require that around every x there exists an open neighbourhood $U \subset M$ and a diffeomorphism $\varphi : \pi^{-1}(U) \to U \times E_x$ such that its projection to the first factor gives the bundle projection: $\mathrm{pr}_1 \circ \varphi = \pi$; φ is called *local trivialization* of the vector bundle. A trivialization of a vector bundle over its whole base is called a *global trivialization*.

A smooth map $f : E \to F$ between two $\mathbb{K}$-vector bundles $\pi_M : E \to M$ and $\pi_N : F \to N$ is a *(vector) bundle homomorphism* if there exists a smooth map $g : M \to N$

---

[1]$\mathbb{K}$ will always be either $\mathbb{R}$ or $\mathbb{C}$. In particular, $\mathbb{K}$ is a field of characteristic 0.

such that $\pi_N \circ f = g \circ \pi_M$ and the restriction of the map to each fibre $f\restriction_{E_x} : E_x \to F_{g(x)}$ is $\mathbb{K}$-linear. In other words we require that the diagram

$$
\begin{array}{ccc}
E & \xrightarrow{\;\;f\;\;} & F \\
\pi_M \downarrow & & \downarrow \pi_N \\
M & \xrightarrow{\;\;g\;\;} & N
\end{array}
$$

commutes and that f is $\mathbb{K}$-linear map on each fibre. By definition, if f is a bundle homomorphism, then g is given by $g = \pi_N \circ f \circ \pi_M^{-1}$.

Given an open subset $U \subset M$, we can restrict $\pi : E \to M$ to a vector bundle $\pi_U : E_U \to U$ by setting $E_U \doteq \pi^{-1}(U)$ and $\pi_U \doteq \pi\restriction_U$. More generally, a subset $E' \subset E$ such that $\pi\restriction_{E'} : E' \to M$ is a vector bundle and $E' \cap \pi^{-1}(x)$ is a vector subspace in $\pi^{-1}(x)$ for all $x \in M$ is called a *vector subbundle* of E. If each fibre of E' has dimension k, we say that E' is a rank-k subbundle of E. For a subbundle $\pi' : E' \to M$ of E we define the *quotient bundle* E/E' as the disjoint union $\bigsqcup_{x \in M} E_x / E'_x$ of the quotient spaces of the fibres.

If M, N are smooth manifolds with a smooth map $\psi : M \to N$ and a vector-bundle $E \to N$, we can define on M the *pullback bundle* $\psi^* E$ as the bundle whose fibres over M are given by $(\psi^* E)_x \doteq E_{\psi(x)}$ for each $x \in M$.

A *section* of a vector bundle E is a continuous map $f : M \to E$ such that $\pi \circ f = \mathrm{id}_M$ (Fig. 1.3); the space of sections of a vector bundle E is denoted by $\Gamma(E)$. We denote by $\Gamma^n(E)$ the space of C^n sections (C^n maps $f : M \to E$) of the vector bundle E, while the spaces of compactly supported sections are indicated by a subscript 0, *e.g.*, $\Gamma_0^\infty(E)$. Furthermore, a *local section* over an open subset $U \subset M$ is a section of the vector bundle E_U.

If N is another smooth manifold with a vector-bundle $F \to N$ and there exists a smooth map $\psi : M \to N$, the *pullback section* $\psi^* f \in \psi^* F$ of $f \in \Gamma(F)$ is defined as the section $\psi^* f \doteq f \circ \psi$. The opposite of the pullback is achieved by the pushforward if ψ is a diffeomorphism. Namely, we will say that $\psi_* h \doteq h \circ \psi^{-1}$ is the *pushforward*
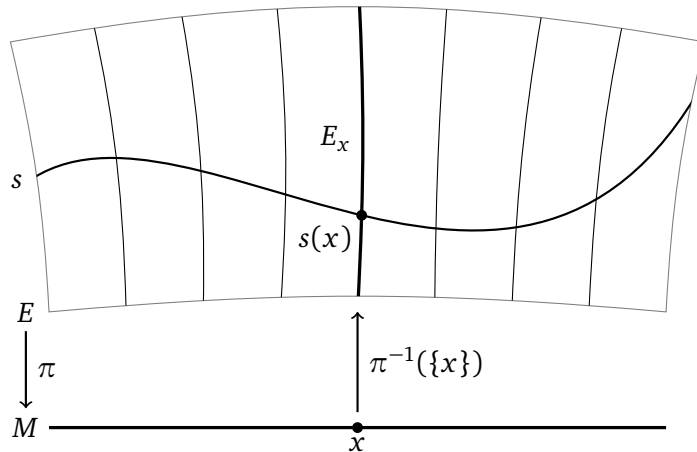

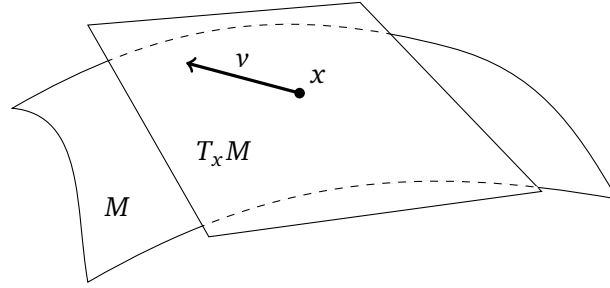**Figure 1.3.** A vector bundle and a section.

Figure 1.4. The tangent space of a manifold at a point and a tangent vector.

section of $h \in \Gamma(E)$. If h is compactly supported in a region $U \subset M$, it suffices that ψ is a diffeomorphism onto the image of U; outside of $\psi(U)$ we set $\psi_* h$ identical to zero.

1.1.3 Tangent bundle

Let M be a smooth manifold. A *(tangent) vector at* $x \in M$ is a linear map $v : C^\infty(M) \to \mathbb{R}$ that satisfies the Leibniz 'product' rule

$$v(f g) = f(x) v g + g(x) v f$$

for all $f, g \in C^\infty(M)$. The set of all tangent vectors constitutes a vector space $T_x M$ called the *tangent space to M at x*; it has the same dimension as the base manifold M for each $x \in M$. Note that, if a smooth manifold M is also a vector space, then we can identify it with its tangent space, *i.e.*, $M \simeq T_x M$ at each point x, which justifies to the geometric visualization of the tangent space (Fig. 1.4).

The vector bundle $TM \to M$ with fibres $T_x M$ at x is the *tangent bundle of M*. In the special situation where the n-dimensional smooth manifold M can be covered by a single chart, TM is diffeomorphic to $M \times \mathbb{R}^n$.

The sections $\Gamma(TM)$ of the vector bundle are called *vector fields*. Applying a vector field $v \in \Gamma(TM)$ to a function $f \in C^\infty(M)$, we obtain a new function $(vf)(x) \mapsto v_x f$, *viz.*, a vector field defines a linear automorphism on the smooth functions called a *derivation*. This is exactly the *Lie derivative* $\mathscr{L}_v f$ of a function $f \in C^\infty(M)$ along a vector field v: $(\mathscr{L}_v f)(x) = (vf)(x)$. The Lie derivative $\mathscr{L}_v w$ of a differentiable vector field w with respect to another differentiable vector field v is another vector field such that

$$\mathscr{L}_v w \doteq [v, w] \doteq v \circ w - w \circ v$$

when applied to smooth functions. It satisfies the Leibniz rule and the Jacobi identity

$$\mathscr{L}_v(f w) = (\mathscr{L}_v f) w + f \mathscr{L}_v w, \quad \mathscr{L}_u[v, w] = [\mathscr{L}_u v, w] + [v, \mathscr{L}_u w]$$

for all $f \in C^\infty(M)$ and vector fields $u, v, w \in \Gamma^\infty(TM)$.

Given two smooth manifolds M, N and a smooth map $F : M \to N$, we can define at each point $x \in M$ the *tangent map* or *differential of F at x* as the linear map
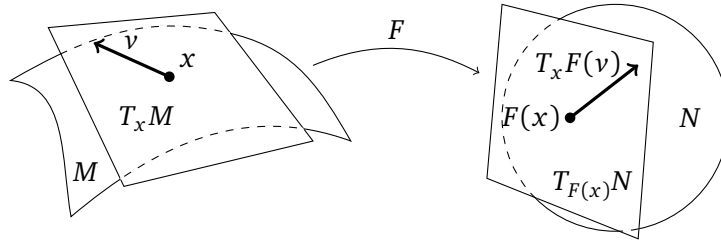
$$T_x F : T_x M \to T_{F(x)} N,$$

Figure 1.5. The tangent map at a point.

see also Fig. 1.5, which is for every $v \in T_x M$ and $f \in C^\infty(M)$ the derivation

$$T_x F(v)(f) = v(f \circ F).$$

The tangent maps of F at all points taken together form $TF : TM \to TN$, the *(global) tangent map* or *differential of F*. If $N = \mathbb{K}$, *i.e.*, F is a smooth function on a manifold, we note that $dF \doteq TF$ is the usual differential.

If F is even a diffeomorphism, then TF defines a bijection between the vector fields on M and N. In this case one can define the *pushforward $F_* v$* of a vector field v on M by F as the vector field on N given at each $x \in N$ by

$$(F_* v)_x \doteq T_{F^{-1}(x)} F\big(v_{F^{-1}(x)}\big).$$

Clearly, this is generally not well-defined if F is not a diffeomorphism. Using that F is invertible, a *pullback $F^* w$* of a vector field w on N can be defined as the inverse of the pushforward, namely,

$$F^* w \doteq (F^{-1})_* w.$$

The tangent map allows us to single out an important type of maps between manifolds: *Immersions* are maps $F : M \to N$ such that TF is injective; if in addition F is injective and a homeomorphisms onto its image, then it is called an *embedding*. Consequently we say that a subset $S \subset M$ is an *immersed submanifold* if it is a topological manifold and the inclusion $S \hookrightarrow M$ is an immersion; if the topology of S is the subspace topology and the inclusion is an embedding, then S is called an *embedded submanifold*. Observe that the tangent space $T_x S$ of a submanifold $S \subset M$ is a subspace of $T_x M$ at every point $x \in S$. We can then say that a vector field v is *tangent to S* if $v_x \in T_x S$ at every point $x \in S$.

1.1.4   Curves

A *parametrized curve* on a smooth manifold M is a map $\gamma : I \to M$ from a connected, usually open, interval $I \subset R$ into the manifold.

A curve $\gamma : [a, b) \to M$ is called *inextendible* if there exists a sequence t_n converging to b such that $\gamma(t_n)$ does not converge. This notion readily extends to left-open domains and open domains.

The *velocity $\dot{\gamma}(t)$* at t of a differentiable parametrized curve $\gamma : I \to M$, where

$I \subset \mathbb{R}$ is an interval, is the vector

$$\dot{\gamma}(t) \doteq T_t \gamma \left(\frac{\mathrm{d}}{\mathrm{d}t} \bigg|_t \right) \in T_{\gamma(t)} M.$$

Working in the opposite direction, we can try to find a curve γ, whose velocity at every point is determined by a given vector field v:

$$\dot{\gamma}(t) = v_{\gamma(t)}. \tag{1.1}$$

Such a curve γ is called an *integral curve* of v.

For a sufficiently small interval I around 0 a unique integral curve starting at a point $x \in M$ can always be found by solving the differential equation (1.1) in a coordinate neighbourhood of x. The domain of an integral curve cannot necessarily be extended to the entire real line. We say that a vector field is *complete* if the domain of all of its *maximal integral curves*, *i.e.*, the integral curves whose domain cannot be extended, is the entire real line.

The *flow ψ_t* of a complete vector field v is

$$\psi_t(x) = \gamma_x(t),$$

where γ_x is the maximal integral curve starting at $x \in M$. This defines for every t a diffeomorphism $\psi_t : M \to M$ and the collection of all these diffeomorphisms is a group $\{\psi_t\}_{t \in \mathbb{R}}$ with unit ψ_0, multiplication $\psi_s \circ \psi_t = \psi_{s+t}$ and inverse $\psi_t^{-1} = \psi_{-t}$. If v is not complete, then one can still define a *local flow ψ_t* around a point x with domain U, where U is a neighbourhood of x, and $t \in I$ is restricted to an interval around 0.

1.1.5 Cotangent bundle

At each point x of a smooth manifold M, we define the *cotangent space to M at x*, denoted $T_x^* M$, as the dual space of tangent space at the same point, namely,

$$T_x^* M \doteq (T_x M)^*.$$

The elements of the cotangent space $T_x^* M$ are called *(tangent) covectors*; naturally they are linear functionals on the tangent space. Taking the cotangent spaces at each point as fibres, we obtain $T^* M$, the *cotangent bundle of M*. The sections of the cotangent bundle are called *covector fields* or *one-forms* (see Sect. 1.3). The differentials (at a point) of functions discussed above are examples of covectors resp. covector fields.

A smooth map $F : M \to N$ between smooth manifolds M, N induces at each point $x \in M$ the map $T_x F : T_x M \to T_{F(x)} N$ between the tangent spaces. By duality one can find the transpose map $T_x^* F : T_{F(x)}^* N \to T_x^* M$, called the *cotangent map of F at x*, between the cotangent spaces at $F(x)$ and x, which is given for each $v \in T_x M$ and $\omega \in T_{F(x)}^* N$ as

$$\left(T_x^* F(\omega) \right)(v) = \omega \left(T_x F(v) \right).$$

This map then gives the *(global) cotangent map* $T^*F : T^*N \to T^*M$ and thus facilitates the definition of the *pullback $F^*\omega$* of a covector field ω on N by F as the covector field on M given at each $x \in M$ by

$$(F^*\omega)_x = T_x^*F\big(\omega_{F(x)}\big).$$

Note that, different than the pushforward, the pullback is even defined if F is not a diffeomorphism. However, if F is a diffeomorphism, we can define the *pushforward* of a covector field as the pullback via the inverse F^{-1}. That is,

$$F_*\eta \doteq (F^{-1})^*\eta,$$

where η is a covector field on M. The pushforward of a compactly supported covector field can also be defined if F is an embedding *and* $\dim M = \dim N$ by setting the $F_*\eta$ to zero outside the image of F.

1.1.6 Tensors product bundles

Given two $\mathbb{K}$-vector bundles $\pi : E \to M$ and $\rho : F \to M$ over the same smooth manifold M, we can define the *tensor product bundle* $\pi \otimes \rho : E \otimes F \to M$, which is just the fibrewise tensor product of vector spaces. Namely, the fibre of $E \otimes F$ at $x \in M$ is $(E \otimes F)_x = (\pi \otimes \rho)^{-1}(\{x\}) = E_x \otimes F_x$.

In particular, we denote by $T_q^p(E)$ the *tensor bundle of type (p,q)* of a vector bundle E:

$$T_q^p(E) \doteq E^{\otimes p} \otimes (E^*)^{\otimes q};$$

the tensor bundle of the (co)tangent bundle is simply denoted by $T_q^p M$ instead of $T_q^p(TM)$. The sections of the tensor product bundle $T_q^p M$ are called *tensor (fields)*, if $p = 0$, we say that a tensor is *covariant*, while we say that it is *contravariant* if $q = 0$.

The Lie derivative defined in Sect. 1.1.3 generalizes to covariant tensor fields by duality in the following way: Given a covariant tensor field $S \in \Gamma^\infty(T_q^0 M)$ and vector fields $v, w_1, \ldots, w_q \in \Gamma^\infty(TM)$, the $\mathscr{L}_v S$ of S along v can be defined by

$$(\mathscr{L}_v S)(w_1, \ldots, w_q) \doteq v\big(S(w_1, \ldots, w_q)\big) - S(\mathscr{L}_v w_1, w_2, \ldots, w_q)$$
$$- \cdots - S(w_1, \ldots, w_{q-1}, \mathscr{L}_v w_q).$$

Note that it satisfies the Leibniz rule

$$\mathscr{L}_v(S \otimes T) = (\mathscr{L}_v S) \otimes T + S \otimes \mathscr{L}_v T,$$

where T is any other differentiable covariant tensor field. Since the Lie derivative on covariant tensor fields defines a Lie derivative on covector fields, this may be used to define a Lie derivative on mixed tensor field.

The pullback or pushforward of a mixed tensor field $S \in \Gamma(T_q^p N)$ or $T \in \Gamma(T_q^p M)$ by a diffeomorphism $F : M \to N$, is defined as the $(p + q)$-fold tensor product of the pullback or pushforward map for (co)vector fields. If S is covariant (*i.e.*, $p = 0$), then the pullback is also defined if F is not a diffeomorphism.

If $F = E$, we can define the *symmetric* and *antisymmetric tensor product bundle* $E \odot E$ and $E \wedge E$ as the quotient bundles under the fibrewise equivalence relations $v \otimes w \sim \pm v \otimes w$ for all $v, w \in E_x$. More generally, we denote by $S^p(E)$ and $\bigwedge^p(E)$ the p-th (anti)symmetric tensor product bundle which satisfy fibrewise the relation

$$v(x_1, \ldots, x_p) = v(x_{\sigma(1)}, \ldots, x_{\sigma(p)}) \qquad \text{or}$$
$$v(x_1, \ldots, x_p) = (\text{sgn}\, \sigma) v(x_{\sigma(1)}, \ldots, x_{\sigma(p)})$$

for $v \in E_x^{\otimes p}$ and all $\sigma \in \mathfrak{S}_p$, the symmetric group of p elements, *cf.* Sect. 4.1.1. That is, the fibres of $S^p(E)$ and $\bigwedge^p(E)$ are the p-th symmetric (resp. exterior) power of the fibres of E. Maps $\text{Sym} : T_0^p(E) \to S^p(E)$ and $\text{Alt} : T_0^p(E) \to \bigwedge^p(E)$ extend from the fibrewise maps

$$\text{Sym}\left(v(x_1, \ldots, x_p)\right) = \frac{1}{p!} \sum_{\sigma \in \mathfrak{S}_p} v(x_{\sigma(1)}, \ldots, x_{\sigma(p)}) \qquad \text{and}$$

$$\text{Alt}\left(v(x_1, \ldots, x_p)\right) = \frac{1}{p!} \sum_{\sigma \in \mathfrak{S}_p} (\text{sgn}\, \sigma) v(x_{\sigma(1)}, \ldots, x_{\sigma(p)}),$$

where $v \in E_x^{\otimes p}$. Moreover, fibrewise products $\odot : S^p(E)_x \times S^q(E)_x \to S^{p+q}(E)_x$ and $\wedge : \bigwedge^p(E)_x \times \bigwedge^q(E)_x \to \bigwedge^{p+q}(E)_x$ are defined as

$$v \odot w = \text{Sym}(v \otimes w) \quad \text{and} \quad v \wedge w = \frac{(p+q)!}{p!\, q!} \text{Alt}(v \otimes w).$$

Another possibility to combine two vector bundles E and F is the *exterior tensor product* $E \boxtimes F \to M \times M$. It is defined as the vector bundle over $M \times M$ with fibre

$$(E \boxtimes F)_x = \pi^{-1}(\{x\}) \otimes \rho^{-1}(\{x'\}) = E_x \otimes F_{x'}$$

over the point (x, x'). The sections of the exterior tensor product $T_q^p M \boxtimes T_s^r M$ are called *bitensor (fields)*.

1.1.7 Metrics

Every vector bundle E has a *dual bundle* $\pi^* : E^* \to M$ which has as its fibre E_x^*, the dual vector spaces of the fibres E_x. We call the natural pairing $f(v)$ of an element $f \in E_x^*$ in the dual fibre on an element $v \in E_x$ in the corresponding fibre a *contraction*.

A canonical isomorphism between E and E^* can be constructed if E carries a *(bundle) metric*, *i.e.*, a map

$$(\cdot, \cdot) : E \times_M E \to \mathbb{K}$$

such that the restriction $(\cdot, \cdot)_x$ to each fibre E_x is a fibrewise non-degenerate bilinear form; a (positive-definite) bundle metric can always be constructed. In other words, a metric on E is a section in the tensor product bundle $\Gamma(E^* \otimes E^*)$. Thus a metric induces a *metric contraction* between two elements of the same fibre E_x.

The *dual metric* $(\cdot, \cdot)^*$ is the unique metric on E^* such that

$$(\omega, \eta)^* = (v, w) \quad \text{with} \quad \omega = (v, \cdot), \eta = (w, \cdot)$$

for all $v, w \in \Gamma(E)$. Moreover, given metrics $(\cdot, \cdot)^E$ and $(\cdot, \cdot)^F$ on E and F, they induce metrics on the tensor product bundle $E \otimes F$ and the exterior tensor product bundle $E \boxtimes F$. For the tensor product bundle it is defined fibrewise for all $v \in E_x, w \in F_x$ by

$$(v \otimes w, v \otimes w)_x^{E \otimes F} = (v, v)_x^E (w, w)_x^F$$

and can be extended to arbitrary pairings by polarization and linearity; an analogous construction works for the exterior tensor product.

In pseudo-Riemannian geometry we find the tangent bundle equipped with a continuous symmetric metric usually denoted

$$g : TM \times_M TM \to \mathbb{R}$$

with dual metric g^* on the cotangent bundle. The canonical isomorphism induced by g between TM and T^*M is given by the *musical isomorphisms 'flat'* $\flat : TM \to T^*M$ and its inverse *'sharp'* $\sharp : T^*M \to TM$:

$$v^\flat \doteq g(v, \cdot), \qquad \omega^\sharp \doteq g(\omega, \cdot).$$

A tuple (M, g) of a smooth manifold M with a metric g on its tangent bundle is called a *pseudo-Riemannian manifold*.

The maximal dimension of subspaces of $T_x M$ where g_x is negative-definite is called the *index* $\mathrm{Ind}(g)$ of g; since g is continuous and non-degenerate, the index constant over the manifold. We distinguish in particular two cases:

(a) If $\mathrm{Ind}(g) = 0$ or, in other words, g is pointwise positive-definite, we say that g is a *Riemannian metric*.

(b) If $\mathrm{Ind}(g) = 1$ (and the manifold at least two-dimensional), we say that g is a *Lorentzian metric*.[2]

We say that (M, g) is a *Riemannian* (*Lorentzian*) *manifold* if the metric g is Riemannian (Lorentzian).

The two prototypical examples for a Riemannian and a Lorentzian manifold are *Euclidean space* and *Minkowski space(time)*: n-dimensional Euclidean space is the smooth manifold over $\mathbb{R}^n$ with a the global chart $(\mathbb{R}^n, \mathrm{id})$, coordinate functions $(x^1, \ldots, x^n)$ and with the *Euclidean metric*

$$\delta \doteq \sum_{i=1}^n \mathrm{d}x^i \otimes \mathrm{d}x^i.$$

In the conventions chosen here, $(1+n)$-dimensional Minkowski spacetime is a smooth manifold modelled on $\mathbb{R}^{1+n}$ with the single chart $(\mathbb{R}^{1+n}, \mathrm{id})$, coordinate functions $(t, x^1, \ldots, x^n)$ and with the *Minkowski metric*

$$\eta \doteq -\mathrm{d}t \otimes \mathrm{d}t + \sum_{i=1}^n \mathrm{d}x^i \otimes \mathrm{d}x^i \tag{1.2}$$

---

[2]This *choice* corresponds to the $-+++$ convention, which we will use. In particle physics one often adopts the opposite convention $+---$, *i.e.*, (M, g) is Lorentzian if $\mathrm{Ind}(g) = n - 1$.

in the coordinate frame. Occasionally one uses the shorthand $\mathbb{M}$ to denote four-dimensional Minkowski space $(\mathbb{R}^4, \eta)$.

Let (M, g_M) and (N, g_N) be pseudo-Riemannian manifolds and $\psi : M \to N$ a diffeomorphism. The map ψ is called a *conformal isometry* if

$$\psi^* g_N = \Omega^2 g_M$$

for some positive function $\Omega \in C^\infty_+(M)$. Conformal isometries preserve the angles between vectors; in particular $g_N(v, w)_{\psi(x)} \gtreqqless 0$ implies $g_M(\psi^* v, \psi^* w)_{(}x) \gtreqqless 0$ for all $v, w \in \Gamma(TN)$ and $x \in M$. In the special case that $\Omega \equiv 1$, we say that ψ is an *isometry*. These notions generalize straightforwardly to immersions and embeddings; and we call these maps *conformal immersions/embeddings* and *isometric immersions/embeddings*.

1.1.8   Frames

A *local frame* of a vector bundle E on M is a set $\{e_\mu\}$ of smooth local sections $\Gamma^\infty(E_U)$ on a domain $U \subset M$ contained in a neighbourhood of a local trivialization such that $\{e_\mu(x)\}$ forms a basis for each fibre E_x over U. The *dual frame* is the set $\{e^\mu\}$ in $\Gamma^\infty(E^*_U)$ that satisfies $e^\mu(e_\nu) = \delta^\mu_\nu$, i.e., the e^μ are the dual basis to e_ν. Naturally, given frames $\{e_\mu(x)\}$ and $\{f_\nu(x)\}$ of vector bundles E, F on M, they induce frames $\{e_\mu(x) \otimes f_\nu(x)\}$ on the tensor product bundle $E \otimes F$; the generalization to the exterior tensor product bundle $E \boxtimes F$ is immediate.

If E is equipped with a metric $(\cdot, \cdot)$, we say that the frame $\{e_\mu\}$ is *orthogonal* ($\mathbb{K} = \mathbb{R}$) or *unitary* ($\mathbb{K} = \mathbb{C}$) if it forms an orthonormal basis for each fibre over U, i.e., if

$$(e_\mu, e_\nu) = \delta_{\mu\nu} \varepsilon_\mu \quad \text{with} \quad \varepsilon_\mu \doteq (e_\mu, e_\mu) = \pm 1.$$

Frames allow us to perform calculations in component form, *viz.*, given a section $s \in \Gamma(E)$ its *components* in the frame $\{e_\mu\}$ on U are given via the dual frame $\{e^\mu\}$ as $s^\mu = s(e^\mu)$ such that $s = s^\mu e_\mu$. This is the first instance where we used the *summation convention*: Unless otherwise noted, summation over *balanced* indices (one upper and one lower) is always implied.

If s is the section of a (exterior) tensor product bundle and frames on the single bundles are given, we use multiple indices to denote the sections. For example, if $g \in \Gamma(T^*M \otimes T^*M)$, we can write

$$g = g_{\mu\nu}\, \mathrm{d}x^\mu \otimes \mathrm{d}x^\nu$$

in terms of the coordinate covectors.

The atlas of a manifold induces natural local frames on the tangent and the cotangent bundle. If (U, φ) is a smooth chart on M in a neighbourhood of x, then the *coordinate vectors*[3]

$$\partial_\mu|_x \doteq (T_x \varphi)^{-1} \left(\partial_\mu|_{\varphi(x)} \right)$$

---

[3]Often we will use the shorthand ∂_μ for $\partial/\partial x^\mu$.

define a basis on $T_x M$ because $T_x \varphi : T_x M \to T_{\varphi(x)} \mathbb{R}^n \simeq \mathbb{R}^n$ is an isomorphism. Together with the coordinate functions these coordinate vectors at every point induce natural coordinates on $(TM)_U$. A local frame for $(T^*M)_U$ is then simply the dual frame $\{dx^\mu\}$.

Note that the coordinate vector fields $\{\partial_\mu\}$ associated to a coordinate chart $\{x^\mu\}$ form a very special local frame of the tangent bundle. Whereas the commutator $[\partial_\mu, \partial_\nu]$ vanishes, this is no longer true in every frame $\{e_\mu\}$, where

$$[e_\mu, e_\nu] = c^\rho{}_{\mu\nu} e_\rho$$

has in general non-vanishing *commutation coefficients* $c^\rho{}_{\mu\nu}$.

1.1.9   Differential operators

Given a vector bundle $E \to M$ of n-dimensional smooth manifold M, a *linear differential operator of order m* (with smooth coefficients) is a linear map $P : \Gamma^\infty(E) \to \Gamma^\infty(E)$ which, in local coordinate $\{x^\mu\}$ on U, is given by

$$P\!\restriction_U = \sum_{|\alpha| \le m} a_\alpha(x) \partial^\alpha,$$

where $\alpha = (\alpha_1, \dots, \alpha_n)$ are multi-indices with $\partial^\alpha = \partial_1^{\alpha_1} \cdots \partial_n^{\alpha_n}$ and the coefficients $a_\alpha : \Gamma^\infty(E) \to \Gamma^\infty(E)$ are linear maps.[4] That is, P is locally defined as a polynomial in the partial derivatives $\{\partial_\mu\}$.

The polynomial

$$p(x, \xi) = \sum_{|\alpha| \le m} a_\alpha(x) \xi^\alpha,$$

where $\xi^\alpha = \xi_1^{\alpha_1} \cdots \xi_n^{\alpha_n}$ and ξ is a covector field with components $\xi = \xi_\mu dx^\mu$, is called the *total symbol* of P. The leading term of $p(x, \xi)$,

$$\sigma_P(x, \xi) = \sum_{|\alpha| = m} a_\alpha(x) \xi^\alpha,$$

is the *principal part* or *principal symbol* of P. While this is not true for the total symbol, one can check that the principal symbol is covariantly defined as a function on the cotangent bundle: $\sigma_P : \Gamma^\infty(S^m(T^*M) \otimes E) \to \Gamma^\infty(E)$.

Suppose that (M, g) is a pseudo-Riemannian manifold. If the principal symbol σ_P of a differential operator P is given by the metric

$$\sigma_P(x, \xi) = -g_x(\xi, \xi) \, \mathrm{id}_{E_x},$$

we say that P is *normally hyperbolic* or, alternatively, that it is a *wave operator*. Normally hyperbolic operators on globally hyperbolic spacetime have a well-posed Cauchy problem and therefore play an important role in quantum field theory on curved spacetime.

Given a second differential operator Q on the same vector bundle E, the composition $P \circ Q$ is also a differential operator and the principal symbol of the composed

---

[4]Although we will not explicitly state this, sometimes we will use differential operators with non-smooth coefficients. In that case the coefficients map into C^k sections.

operator is given by the composition of the principal symbols: $\sigma_{P \circ Q} = \sigma_P \circ \sigma_Q$. We say that P is *pre-normally hyperbolic* if there exists Q such that $P \circ Q$ is normally hyperbolic. One can show that also $Q \circ P$ is normally hyperbolic and thus Q is pre-normally hyperbolic too [160].

1.1.10   Index notation

We just saw that in a frame $\{e_\mu\}$ of a vector bundle $E \to M$ we can calculate the components of sections with respect to that frame. For example, given two vectors fields $v, w \in \Gamma(TM)$ on a pseudo-Riemannian manifold (M, g), we can write their metric contraction in terms of their components with respect to the coordinate (co)vector fields $\{\partial_\mu\}$ and $\{\mathrm{d}x^\mu\}$ in a coordinate neighbourhood:

$$g(v, w) = g_{\mu\nu} v^\mu w^\nu.$$

Repeated indices imply summation by the Einstein summation convention as usual.

Usually the frame is not explicitly mentioned but instead implicitly given by a selection of letters for the indices. Henceforth the small Greek letters $\mu, \nu, \lambda, \rho, \sigma$ will always be indices for a coordinate frame of the (co)tangent bundle in the *concrete index notation*.

When calculating contractions between more complicated tensors the notation in terms of indices is often over the abstract index-free notation which quickly becomes unwieldy. Moreover, if the horizontal position of indices is kept fixed, we can use a metric to lower and raise indices, *e.g.*, returning to our example, we write

$$g(v, w) = v_\mu w^\mu.$$

That is, we identify the components of the vector field v with the components of the associated covector field $g(v, \cdot)$. Contracting a tensor $S \in \Gamma(T^2_1 M)$ with the vector fields v, w, we see the advantage of this notation

$$S\big(g(v, \cdot), g(w, \cdot)\big) = g_{\nu\rho} g_{\lambda\sigma} S_\mu{}^{\rho\sigma} v^\mu v^\nu w^\lambda = S_{\mu\nu\lambda} v^\mu v^\nu w^\lambda$$

The 'natural' position of the indices of the tensor S must, however, be agreed upon beforehand.

Note that the formal aspects of this notation do not necessitate the existence of a frame. This leads to Penrose's *abstract index notation*. Even in the absence of a concrete frame, we write for example

$$S\big(g(v, \cdot), g(w, \cdot)\big) = g_{bd} g_{ce} S_a{}^{de} v^a v^b w^c = S_{abc} v^a v^b w^c$$

Now, an index only labels a slot in the index-free expression and does not carry any numerical value. In particular, Einstein summation convention does not apply to abstract indices – it would not even make sense – and double indices only imply (metric) contractions. We will often use the small Latin letters a, b, c, d, e as abstract indices for the (co)tangent bundle.

Both for abstract and concrete index notation it is useful to introduce some shorthands. Symmetrization and antisymmetrization of tensors are denoted by

parentheses and brackets:

$$S_{(ab)} \doteq \tfrac{1}{2}(S_{ab} + S_{ba}), \quad S_{[ab]} \doteq \tfrac{1}{2}(S_{ab} - S_{ba})$$

with the obvious generalization to higher-order tensors. Partial and covariant derivatives (see below) are sometimes indicated by comma and semicolon:

$$(\cdots)_{,a} \doteq \partial_a(\cdots), \quad (\cdots)_{;a} \doteq \nabla_a(\cdots).$$

1.2   Connections and curvature

A (Koszul) *connection* ∇^E on a m-dimensional $\mathbb{K}$-vector bundle $E \to M$ is a $\mathbb{K}$-linear map $\nabla^E : \Gamma^\infty(E) \to \Gamma^\infty(E \otimes T^*M)$ that satisfies the Leibniz rule

$$\nabla^E(\varphi f) = \varphi \nabla^E f + f \otimes d\varphi$$

for all $\varphi \in C^\infty(M)$ and $f \in \Gamma^\infty(E)$. Every vector bundle admits a connection. Henceforth we will often drop the superscript indicating the vector bundle that the connection acts on and simply denote it by ∇. Given a vector field v, the connection ∇ defines the *covariant derivative along v* as $\nabla_v : \Gamma^1(E) \to \Gamma(E)$ with

$$\nabla_v \cdot \doteq (\nabla \cdot)(v).$$

If, in addition, the vector bundle E is equipped with a C^∞ bundle metric $(\cdot, \cdot)$, we say that ∇ is a *metric connection* if

$$v(f, h) = (\nabla_v f, h) + (f, \nabla_v h)$$

holds for all $f, h \in \Gamma^1(E)$. A connection ∇ on $E = TM$ is *torsion-free* if the Lie bracket of two vector fields v, w is given by $[v, w] = \nabla_v w - \nabla_w v$.

Let E, F be two vector bundles with connections ∇^E, ∇^F and sections $f \in \Gamma^1(E)$, $h \in \Gamma^1(F)$ and $u \in \Gamma^1(E^*)$. A connection on the the tensor product bundle $E \otimes F$ is defined by

$$\nabla_v^{E \otimes F}(f \otimes h) = (\nabla_v^E f) \otimes h + f \otimes (\nabla_v^F h).$$

Moreover, a *dual connection* is obtained from

$$(\nabla_v^{E^*} u)(f) = v(u(f)) - u(\nabla_v^E f).$$

This can be used to extend a connection ∇ on a vector bundle E to its dual bundle E^* and more generally to the tensor bundle $T_q^p E$.

If $\psi : M \to N$ is a diffeomorphism between two manifolds and $E \to N$ a vector bundle with a connection ∇^E, then we automatically find a unique connection $\psi^*\nabla$ on the pullback bundle ψ^*E. The *pullback connection* is given by

$$(\psi^*\nabla)_v(\psi^*f) \doteq \psi^*(\nabla_{d\psi(v)}^E f)$$

for all $f \in \Gamma^1(E)$ and $v \in \Gamma(TM)$.

1.2.1 Levi-Civita connection

The *Levi-Civita connection* is the unique metric connection ∇ on TM with C^∞ metric g that is torsion free. Therefore it satisfies the *Koszul formula*

$$
\begin{aligned}
2g(\nabla_u v, w) = {}& u\,g(v,w) + v\,g(u,w) - w\,g(u,v) \\
& - g(u,[v,w]) - g(v,[u,w]) - g(w,[u,v]).
\end{aligned}
\tag{1.3}
$$

Usually we will call the covariant derivative associated to the Levi-Civita connection of (M,g) simply *the* covariant derivative.

In a chart (U,φ) of M, we have

$$
\nabla_{\partial_\mu} \partial_\nu = \Gamma^\rho{}_{\mu\nu}\, \partial_\rho
$$

and we call $\Gamma^\rho{}_{\mu\nu}$ the *Christoffel symbols*[5] of the Levi-Civita connection ∇ for the chart (U,φ). The Christoffel symbols are symmetric in their lower indices $\Gamma^\rho{}_{\mu\nu} = \Gamma^\rho{}_{\nu\mu}$ because $[\partial_\nu, \partial_\nu] = 0$. Given two vector fields $v, w \in \Gamma^1(TM)$, written in the coordinate basis as $v = v^\mu \partial_\mu$ and $w = w^\mu \partial_\mu$, w has covariant derivative

$$
\nabla_v w = v^\mu \left(\frac{\partial w^\rho}{\partial x^\mu} + \Gamma^\rho{}_{\mu\nu}\, w^\nu \right) \partial_\rho
$$

along v. The Kozul formula (1.3) yields a formula for the Christoffel symbols

$$
\Gamma^\rho{}_{\mu\nu} = \frac{1}{2} g^{\rho\lambda} \left(\frac{\partial g_{\mu\lambda}}{\partial x^\nu} + \frac{\partial g_{\nu\lambda}}{\partial x^\mu} - \frac{\partial g_{\mu\nu}}{\partial x^\lambda} \right).
$$

1.2.2 Curvature of a connection

Different from ordinary second derivatives, covariant second derivatives

$$
(\nabla^2 f)(v,w) = \nabla_v \nabla_w f - \nabla_{\nabla_v w} f,
$$

where $v, w \in \Gamma^1(TM)$ and $f \in \Gamma^2(E)$, do not commute in general. The *curvature* $F \in \Gamma(T^*M \otimes T^*M \otimes E \otimes E^*)$ of a vector bundle E with connection ∇ quantifies this failure of the covariant second derivative to commute and we define it as

$$
F(v,w) \doteq \nabla_v \nabla_w - \nabla_w \nabla_v - \nabla_{[v,w]}.
$$

We say that the connection is *flat* its curvature F vanishes. If $E = TM$ and ∇ is the Levi-Civita connection of a pseudo-Riemannian manifold (M,g) with smooth metric g, we denote the curvature by $R(v,w)$ instead and call it *Riemann curvature (tensor)* of (M,g).

Let $u, v, w \in \Gamma^2(TM)$ be vector fields and $f \in \Gamma^2(E)$ a section of E. By definition the curvature is skew-hermitian: $F(v,w) = -F(w,v)$. If the connection ∇ is compatible with a metric $(\cdot,\cdot)$ on E, then we also have that F is skew-adjoint, $(F(v,w)f, f) = -(f, F(v,w)f)$, and that it satisfies the *first Bianchi identity*

$$
(\nabla_u F)(v,w) + (\nabla_v F)(w,u) + (\nabla_w F)(u,v) = 0.
$$

---

[5] Note that ∇ is not a tensor and thus the Christoffel symbols do not transform as tensors.
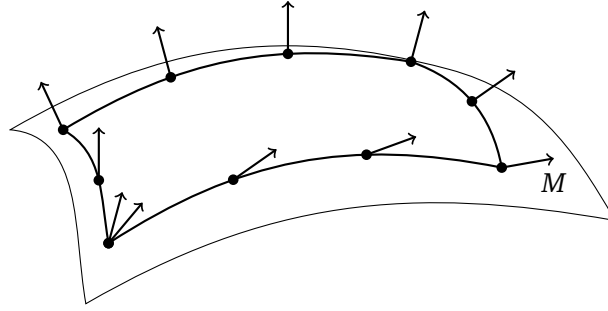
Figure 1.6. The parallel transport of a vector along a closed path.

Furthermore, if $E = TM$ and ∇ the Levi-Civita connection, also the *second Bianchi identity*

$$R(u, v)w + R(v, w)u + R(w, u)v = 0$$

holds.

Several other curvature tensors can be derived from the Riemann curvature. The *Ricci curvature (tensor)* is defined as the symmetric $(0, 2)$ tensor $\mathrm{Ric}(v, w) = \mathrm{tr}(u \to R(v, u)(w))$, where $u, v, w \in \Gamma^2(TM)$; if it vanishes, we say that (M, g) is *Ricci-flat*. Furthermore, we obtain the *Ricci scalar* as the contraction of the Ricci tensor $\mathrm{Scal} \doteq \mathrm{tr}_g \mathrm{Ric} = \mathrm{tr}\,\mathrm{Ric}^\sharp$. Using the coordinate (co)vectors, we can express the Riemann tensor in component form as:

$$R^\sigma{}_{\mu\nu\lambda} = \partial_\nu \Gamma^\sigma{}_{\mu\lambda} - \partial_\lambda \Gamma^\sigma{}_{\mu\nu} + \Gamma^\sigma{}_{\rho\lambda}\,\Gamma^\rho{}_{\mu\nu} - \Gamma^\sigma{}_{\rho\nu}\,\Gamma^\rho{}_{\mu\lambda}.$$

The Ricci tensor and Ricci scalar are then written as $R_{\mu\nu} = -\delta^\lambda_\sigma R^\sigma{}_{\mu\nu\lambda}$ and $R = g^{\mu\nu} R_{\mu\nu}$.

1.2.3   Geodesics

Let $\gamma : I \to M$ be a smooth curve parametrized by an interval $I \subset \mathbb{R}$ and $f \in \Gamma^1(\gamma^* E)$ a section in the pullback of the vector bundle $E \to M$. The *covariant derivative of f along γ* is given by

$$\frac{\nabla f}{\mathrm{d}t} \doteq (\gamma^* \nabla_{\dot\gamma(t)})f,$$

i.e., it is a special case of a pullback connection. If the covariant derivative of f vanishes along γ, we say that f is *parallel* to γ. Therefore connections facilitate the notion of *parallel transport* along curves (Fig. 1.6, see also Sect. 1.4.2). Namely, given a $s_0 \in E_{\gamma(t_0)}$ at the point $\gamma(t_0)$, the parallel transport of f_0 along γ is the unique solution f of the ordinary differential equation $\nabla(\gamma^* f)/\mathrm{d}t = 0$ with initial condition $(f \circ \gamma)(t_0) = f_0$.

Auto-parallel curves, *i.e.*, curves that satisfy

$$\frac{\nabla}{\mathrm{d}t} \gamma^* \dot\gamma(t) = 0,$$

are called *geodesics of the connection* ∇. These are the usual geodesics (local minimizers of arc length if the metric is Riemannian) with respect to a metric g of a
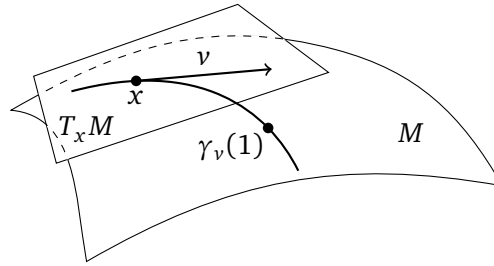
Figure 1.7. The exponential map applied to a vector.

pseudo-Riemannian manifold (M, g) if ∇ is the Levi-Civita connection with respect to g.

It follows from the theory of ordinary differential equations that, given a point $x \in M$ and any vector $v \in T_x M$, there exists a unique geodesic γ_v such that $\gamma_v(0) = x$ and $\dot\gamma_v(0) = v$. Let Υ_x be the set of vectors v at x that give an inextendible geodesic γ_v defined (at least) on the interval $[0, 1]$. We say that M is *geodesically complete* if $\Upsilon_x = T_x M$ at every $x \in M$. Nevertheless, if there is a vector $v \notin \Upsilon_x$, there exists $\varepsilon \in \mathbb{R}$ such that $\varepsilon v \in \Upsilon_x$.

The *exponential map* at x is defined as the map (Fig. 1.7)

$$\exp_x : \Upsilon_x \to M, \quad v \mapsto \gamma_v(1).$$

That is, remembering that geodesics can be linearly reparametrized, the exponential map $\exp_x$ maps vectors at x to geodesics through x. For each $x \in M$ there exists an open neighbourhood $U' \subset T_x M$ of the origin on which $\exp_x$ is a diffeomorphism into an open neighbourhood $U \subset M$ of x. If U' is starshaped[6], then we say that U is *geodesically starshaped* (U is a *normal neighbourhood*) with respect to x. Moreover, if $U \subset M$ is geodesically starshaped with respect to all of its points, it is called *geodesically convex*.

1.2.4   Killing vector fields

Given an n-dimensional pseudo-Riemannian manifold (M, g), a *Killing vector field* is a vector field κ such that

$$\mathscr{L}_\kappa g = 0.$$

In terms of the Levi-Civita connection on (M, g), this equation may also be written as $\nabla_a \kappa_b - \nabla_b \kappa_a = 0$ in the abstract index notation. More generally, a *conformal Killing vector field* is a vector field κ such that

$$\mathscr{L}_\kappa g = \omega g \quad \text{with} \quad \omega = \frac{2}{n} \operatorname{tr}(\nabla \kappa).$$

---

[6]A *starshaped neighbourhood* S of a vector space V is an open neighbourhood $S \subset V$ of the origin such that $tv \in S$ for all $t \in [0, 1]$ and $v \in S$.

An equivalent definition of the Lie derivative of a covariant tensor $S \in \Gamma^1(T^0_q M)$ along a differentiable vector field κ is given by

$$\mathscr{L}_\kappa S = \lim_{t \to 0} \frac{1}{t}\left(\psi_t^* S - S\right)$$

where ψ_t is the (local) flow of κ. Therefore, a vector field κ is (conformal) Killing vector field if and only if the flow that it generates is a family of local (conformal) isometries. In other words, κ encodes a *(conformal) symmetry* of (M, g).

Now, if γ is a geodesic and κ a Killing vector field, then it holds that $g(\kappa, \dot\gamma)$ is constant along γ. That is, the geodesics of (M, g) correspond to conserved quantities under the symmetry given by κ. If κ is a *conformal* Killing vector field, then $g(\kappa, \dot\gamma)$ is constant only if $g(\dot\gamma, \dot\gamma) = 0$.

1.3 Differential forms and integration

As already noted above, sections of the cotangent bundle T^*M are called *(differential) 1-forms*. More generally, sections of the p-th antisymmetric tensor bundle $\bigwedge^p(T^*M)$ are called *(differential) p-forms*. The set of all smooth p-forms on M is usually denoted $\Omega^p(M) \doteq \Gamma^\infty(\bigwedge^p T^*M)$.

1.3.1 Exterior derivative

The *exterior derivative* $\mathrm{d} : \Omega^p(M) \to \Omega^{p+1}(M)$ is the unique generalization of the differential of functions such that:

(a) $\mathrm{d}f$ for 0-forms (*i.e.*, functions) $f \in \Omega^0(M)$ is the usual differential,

(b) d is a $\wedge$-antiderivation, *i.e.*, it satisfies the product rule

$$\mathrm{d}(\omega \wedge \eta) = \mathrm{d}\omega \wedge \eta + (-1)^p \omega \wedge \mathrm{d}\eta,$$

where $\omega \in \Omega^p(M)$ and $\eta \in \Omega^l(M)$,

(c) $\mathrm{d}^2 = \mathrm{d} \circ \mathrm{d} = 0$,

(d) d commutes with pullbacks.

We say that a form ω is *closed* if $\mathrm{d}\omega = 0$ and *exact* if $\omega = \mathrm{d}\eta$ for some form η. While a closed form is in general not exact, the opposite is obviously always true. The extend to which closed forms fail to be exact is measured by the *de Rham cohomology groups* $H^p_{\mathrm{dR}}(M)$ of the smooth manifold M

$$H^p_{\mathrm{dR}}(M) \doteq \frac{\{\omega \in \Omega^p(M) \mid \omega \text{ closed}\}}{\{\omega \in \Omega^p(M) \mid \omega \text{ exact}\}}.$$

Replacing p-forms with compactly supported forms in the definition above, we obtain the related notion of the *de Rham cohomology group with compact support* $H^p_{\mathrm{dR},0}(M)$.

It is not difficult to show that the de Rham cohomology is a homotopy invariant and thus a topological invariant. This is quite astonishing considering that its definition relies on the smooth structure of the manifold. Note that, if M is *contractible*, *i.e.*, homotopy equivalent to a point, then all its de Rham cohomology groups vanish.

1.3.2 Integration

A smooth n-form $\mu \in \Omega^n(M)$ on a smooth n-dimensional manifold M is called a *volume form* if $\mu(x) \neq 0$ for all $x \in M$. If such a volume form μ exists, we say that M is *orientable* because μ assigns a consistent *orientation* to all of M. Here, we say that a basis $v_1, \ldots, v_n \in T_x M$ is *positively (negatively) oriented* at x with respect to $\omega \in \Omega^n(M)$ if $\omega(x)(v_1 \otimes \ldots \otimes v_n) \gtrless 0$. If M is an orientable smooth manifold equipped with a metric tensor g, there exists a unique volume form μ_g, the *g-volume*, which satisfies

$$\mu_g(x)(v_1 \otimes \cdots \otimes v_n) = \sqrt{|g|} \doteq \sqrt{|\det[g_x(v_i, v_j)]|}$$

for a positively oriented basis $v_1, \ldots, v_n \in T_x M$.

Once integration over $\mathbb{R}^n$ is defined, *i.e.*, given a domain $U \subset \mathbb{R}^n$

$$\int_U v = \int_U f \, \mathrm{d}x^1 \wedge \cdots \wedge \mathrm{d}x^n = \int_U f \, \mathrm{d}x^1 \cdots \mathrm{d}x^n$$

for some $f \in C^\infty(\mathbb{R}^n)$ such that $v = f \, \mathrm{d}x^1 \wedge \cdots \wedge \mathrm{d}x^n$, we can use the a partition of unity, local charts and linearity of the integral to extend the notion to general smooth manifold. That is, if $\omega \in \Omega^n(M)$ is a form of maximal degree on an orientable smooth manifold M which is compactly supported in the chart (U, φ), then

$$\int_M \omega = \pm \int_{\varphi(U)} \varphi_* \omega,$$

where the sign depends on the orientation of the chart (U, φ) with respect to ω. Thus we obtained a method of *integration on manifolds* that is invariant under orientation-preserving diffeomorphism invariant.

Integrating a metric $(\cdot, \cdot)$ on a vector bundle E over an orientable pseudo-Riemannian manifold (M, g), yields a natural inner product $\langle \cdot, \cdot \rangle$ on the sections of E: Given $f, h \in \Gamma(E)$, we define

$$\langle f, h \rangle \doteq \int_M (f, h) \, \mu_g,$$

whenever the integral exists.

Arguably the most important result on integration on manifolds is *Stokes' theorem* – a generalization of the fundamental theorem of calculus. It states that the integral of an exact form $\mathrm{d}\omega \in \Omega^n(M)$ over a relatively compact open subset $U \subset M$ of a n-dimensional oriented manifold M is given by the integral of ω over the C^1-boundary ∂U:

$$\int_U \mathrm{d}\omega = \int_{\partial U} \iota^* \omega,$$

where $\iota : \partial U \hookrightarrow M$ denotes the inclusion map. Note that the classical theorems of Gauss (also called the divergence theorem) and Green are special cases of Stokes' theorem.

1.3.3   Hodge star and codifferential

Let (M, g) be an oriented pseudo-Riemannian n-dimensional manifold for the remainder of this section.

We can now introduce a smooth bundle homomorphism $* : \bigwedge^p(T^*M) \to \bigwedge^{n-p}(T^*M)$, the *Hodge star (operator)*; it is the unique bijection such that

$$\omega \wedge * \eta = g(\omega, \eta)\mu_g$$

for all $\omega, \eta \in \Omega^p(M)$. This implies the properties

$$*1 = \mu_g, \quad *\mu_g = (-1)^{\mathrm{Ind}(g)}, \quad **\omega = (-1)^{\mathrm{Ind}(g)+p(n-p)}\omega.$$

It also follows that we can rewrite the inner product $\langle \cdot, \cdot \rangle$ between differential forms induced by the metric as

$$\langle \omega, \eta \rangle = \int_M \omega \wedge * \eta.$$

for all $\omega, \eta \in \Omega^p(M)$ for which the integral is defined.

The formal adjoint of the exterior derivative with respect to this pairing is the *codifferential* $\delta : \Omega^{p+1}(M) \to \Omega^p(M)$:

$$\langle \omega, \delta\eta \rangle \doteq \langle \mathrm{d}\omega, \eta \rangle$$

or, equivalently,

$$\delta\omega \doteq (-1)^{np+1+\mathrm{Ind}(g)} * \mathrm{d} * \omega.$$

The codifferential is *not* a derivation and thus it does not satisfy the Leibniz rule. Note that for one-forms η the codifferential satisfies $\delta\eta = -\operatorname{div}\eta^\sharp$, *i.e.*, it is equal to minus the divergence of the related vector field.

Analogously to the case of the exterior derivative, a form ω is called *coclosed* if $\delta\omega = 0$ and *coexact* if $\omega = \delta\eta$ for some form η. As a consequence of the bijectivity of the Hodge star one finds

$$H_{\mathrm{dR}}^{n-p}(M) \cong \frac{\{\omega \in \Omega^p(M) \mid \omega \text{ coclosed}\}}{\{\omega \in \Omega^p(M) \mid \omega \text{ coexact}\}},$$

and, in particular,

$$H_{\mathrm{dR}}^p(M) \cong H_{\mathrm{dR},0}^{n-p}(M)$$

if $H_{\mathrm{dR}}^p(M)$ is finite-dimensional; the latter relation is a consequence of the *Poincarè duality* theorem.

A normally hyperbolic differential operator generalizing the usual d'Alembert operator, the *Laplace–de Rham operator*, can be obtained by composition of the codifferential and exterior derivative; we define it as

$$\Box \doteq \mathrm{d} \circ \delta + \delta \circ \mathrm{d}.$$

In abstract tensor notation the Laplace–de Rham operator acting on a (smooth) p-form ω is given by the Weitzenböck type formula [145, Eq. (10.2)]

$$(\Box\omega)_{a_1\cdots a_p} = -\nabla^b\nabla_b\omega_{a_1\cdots a_p} + \sum_m R_{a_m b}\omega_{a_1\cdots}{}^b{}_{\cdots a_p} + \sum_{m\neq n} R_{a_m b a_n c}\omega_{a_1\cdots}{}^b{}_{\cdots}{}^c{}_{\cdots a_p}.$$

Consequently the Laplace–de Rham operator on scalar functions (0-forms), which we will also call the *d'Alembert operator*, is defined as

$$\Box \doteq \delta d = -\nabla^a\nabla_a,$$

which differs from the definition in some publications of the author [3, 4] by a sign.

1.3.4 Integral submanifolds

While a vector field can always be locally integrated to give an integral curve, it is not true that to every rank $k > 1$ subbundle of the tangent bundle TM there exists a m-dimensional submanifold of a smooth manifold M.

Let $D \subset TM$ be a smooth rank-k subbundle, a *plane field*, and $N \subset M$ an immersed submanifold. N is an *integral manifold* of D if

$$T_x N = D_x$$

for every $x \in N$ and we say that D is *integrable*.

The plane field D is locally spanned by smooth vector fields $v_1,\ldots,v_k$, i.e., each local frame around a point $x \in M$ is given by the vector fields v_i such that $v_1\restriction_x,\ldots,v_k\restriction_x$ is a basis of D_x. *Frobenius' theorem* states that D is integrable if and only if $[v_i,v_j]_x \in D_x$ for all i,j at every point x.

A plane field D is equivalently locally specified in a neighbourhood $U \subset M$ by a collection of covector fields $\omega_1,\ldots,\omega_{n-k}$ such that

$$D_x = \ker\omega_1\restriction_x \cap \cdots \cap \ker\omega_{n-k}\restriction_x$$

for every $x \in U$. The dual to Frobenius theorem is then: D is integrable if and only if there exist smooth covector fields $\{\eta_{i,j} \mid i,j = 1,\ldots,n-k\}$ such that

$$d\omega_i = \sum_{j=1}^{n-k} \omega_j \wedge \eta_{i,j},$$

in other words, D is *involutive*. For example, a smooth covector field ω specifies a codimension 1 integral manifold of M if and only if

$$d\omega \wedge \omega = 0.$$

A *foliation of dimension k* of a smooth manifold M is a collection $\mathcal{F} = (N_i)$ of disjoint, connected, (non-empty) immersed k-dimensional submanifolds of M, the *leaves of the foliation*, such that their union is the entire manifold M and each point $x \in M$ has a chart (U,φ) with coordinates $(x_1,\ldots,x_n)$ such that for every leaf N_i that intersects U the connected components of the image $\varphi(U \cap N_i)$ are given by the equations $x_{k+1} = \text{const},\ldots,x_n = \text{const}$.

If $\mathcal{F}$ is a foliation of M, then the tangent spaces of the leaves form a plane field of M which is involutive. Conversely, the *global Frobenius theorem* states that the maximal integral manifolds of an involutive plane field D on M form a foliation of M.

1.4   Bitensors

Let (M, g) be a pseudo-Riemannian n-dimensional manifold for the remainder of this section. Henceforth we will use unprimed $a, b, \dots$ and primed $a', b', \dots$ abstract indices to distinguish between objects that transform like tensors at x and x' respectively. For example, a bitensor $T \in \Gamma(T_0^2 M \boxtimes T_1^0 M)$ can be denoted by $T^{ab}{}_{c'}(x, x')$, where relative position of primed and unprimed indices is arbitrary. Note that $T^{ab}{}_{c'}$ transforms like a contravariant 2-tensor at x and as a covector at x'. While many of the operations presented in this section generalize to sections of arbitrary exterior tensor bundles, we limit the discussion to bitensors, *i.e.*, sections of $T_q^p M \boxtimes T_s^r M$, in favour of concreteness rather than generality.

Taking covariant derivatives of bitensors, we further notice that derivatives with respect to x and x' commute with each other. That is, every (sufficiently regular) bitensor $T(x, x')$ satisfies the identity

$$T_{;ab'} = T_{;b'a},$$

where we have suppressed any other indices (we will do the same in the next two paragraphs). Partial derivatives commute as always.

Often one is interested in the limiting behaviour $x' \to x$. This limit is called the *coincidence* or *coinciding point limit* and can be understood as a section of a tensor bundle of M whenever the limit exists and is independent of the path $x' \to x$. If a unique limit exists, we adopt *Synge's bracket notation* [209, Chap. II.2]

$$[T](x) \doteq \lim_{x' \to x} T(x, x').$$

An important result on coincidence limits of bitensors is *Synge's rule*: Let $T(x, x')$ be a bitensor as above, then [179, Chap. I.4.2]

$$[T]_{;a} = [T_{;a}] + [T_{;a'}] \quad \text{or} \quad [T_{;a'}] = [T]_{;a} - [T_{;a}] \tag{1.4}$$

whenever the limits exist and are unique. The second equality is a useful tool to turn unprimed derivatives into primed ones and vice-versa.

1.4.1   Synge's world function

In a geodesically convex neighbourhood U we can define the *geodesic distance* between two points $x, x' \in U$ as the arc length of the unique geodesic γ joining $x = \gamma(t_0)$ and $x' = \gamma(t_1)$. It is given by

$$d(x, x') \doteq \int_{t_0}^{t_1} g_{\gamma(t)}(\dot{\gamma}, \dot{\gamma})^{1/2} \, \mathrm{d}t = (t_1 - t_0) g_x(\dot{\gamma}, \dot{\gamma})^{1/2}$$

because the integrand is constant along the geodesic as a consequence of the geodesic equation. A slightly more useful function is *Synge's world function*, introduced in

[186, 187, 208], which gives half the squared geodesic distance between two points (thus it is sometimes also called the *half, squared geodesic distance*). Namely,

$$\sigma(x,x') \doteq \tfrac{1}{2}d(x,x')^2 = \tfrac{1}{2}(t_1 - t_0)^2 g_x(\dot{\gamma},\dot{\gamma}). \tag{1.5}$$

In terms of the exponential map the geodesic distance and the world function can be expressed as

$$d(x,x') = g_x\big(\exp_x^{-1}(x'), \exp_x^{-1}(x')\big)^{1/2}$$
$$\sigma(x,x') = \tfrac{1}{2}g_x\big(\exp_x^{-1}(x'), \exp_x^{-1}(x')\big).$$

Note that geodesic distance d and Synge's world function σ are invariant under linear reparametrizations of the geodesic γ. Furthermore, they are both examples of bitensors – in fact, *biscalars*.

For the covariant derivatives of the world function it is common to write

$$\sigma_{a_1\cdots a_p b'_1\cdots b'_q} \doteq \nabla_{a_1}\cdots\nabla_{a_p}\nabla_{b'_1}\cdots\nabla_{b'_q}\sigma,$$

i.e., we always omit the semicolon on the left-hand side. From (1.5) one can compute

$$\sigma_a(x,x') = (t_1 - t_0)g_{ab}\dot{\gamma}^b \quad \text{and} \quad \sigma_{a'}(x,x') = (t_0 - t_1)g_{a'b'}\dot{\gamma}^{b'},$$

where the metric and the tangent vector are evaluated at x and x' respectively. Consequently we have $[\sigma_a] = [\sigma_{a'}] = 0$.

According to (1.5), the norm of these covectors is given by the fundamental relation

$$\sigma_a\sigma^a = 2\sigma = \sigma_{a'}\sigma^{a'}. \tag{1.6}$$

Therefore, σ_a and $\sigma_{a'}$ are nothing but tangent vectors at x and x' to the geodesic γ with length equal to the geodesic distance between these points. Actually, (1.6) together with the initial conditions

$$[\sigma] = 0 \quad \text{and} \quad [\sigma_{ab}] = [\sigma_{a'b'}] = g_{ab} \tag{1.7}$$

can be taken as the definition of the world function. Coincidence limits for higher derivatives of σ can be obtained by repeated differentiation of (1.6) in combination with Synge's rule (1.4). In particular one finds [179]

$$[\sigma_{abc}] = [\sigma_{abc'}] = [\sigma_{ab'c'}] = [\sigma_{a'b'c'}] = 0. \tag{1.8}$$

An important biscalar that can be constructed from the world function is the *van Vleck–Morette determinant* defined as [156]

$$\Delta(x,x') \doteq \operatorname{sgn}(\det g_x)(\det g_x \det g_{x'})^{-1/2}\det\big(-\sigma_{ab'}(x,x')\big) > 0. \tag{1.9}$$

The van Vleck–Morette determinant expresses geodesic (de)focusing: $\Delta > 1$ implies that geodesics near x and x' undergo focusing whereas $\Delta < 1$ implies that these geodesics undergo defocusing [179].

From (1.9) one can derive the transport equation

$$n = \sigma^a(\ln\Delta)_{,a} + \sigma^a{}_a = \sigma^{a'}(\ln\Delta)_{,a'} + \sigma^{a'}{}_{a'}.$$

This, together with the initial condition $[\Delta] = 1$, can also be used as an alternative definition of the van Vleck–Morette determinant.

1.4.2 Parallel propagator

Another important biscalar is the *parallel propagator*. It does exactly what its name suggests: in a geodesically convex neighbourhood U it transports a vector of a fibre at a point $x' \in U$ to a vector at $x \in U$ along the unique geodesic γ joining the two points. We will denote it by the same symbol as the metric. Here we define it only for the tangent bundle, where we write

$$v^a \doteq g^a{}_{a'} v^{a'}$$

with $v^a \in T_x M$ and $v^{a'} \in T_{x'} M$.

   The parallel propagator satisfies the transport equation

$$g^a{}_{b';c}\, \sigma^c = 0 = g^a{}_{b';c'}\, \sigma^{c'}$$

because σ^a and $\sigma^{a'}$ are tangent to the geodesic γ at x and x' respectively. Together with the initial condition

$$[g^a{}_{b'}] = g^a{}_b = \delta^a_b \tag{1.10}$$

this equation may be taken as the definition of the parallel propagator. With the parallel propagator we can raise and lower the (un)primed index in $g^a{}_{b'}$ with the usual metric and use it to transport (un)primed indices of tensors to their opposite.

1.4.3 Covariant expansion

It is possible to generalize Taylor's series expansion method to bitensors (and thereby also to tensors). The covariant expansion was originally developed for ordinary tensors [186, 187] but can be easily extended to bitensors, see for example [179, Chap. I.6].

   While Taylor's method is used to expand a function $f(x')$ around a point x in terms of powers of the distance $x' - x$ with series coefficients given by the derivatives of f at x, the covariant expansion method replaces functions with bitensors, distance with geodesic distance as given by $\sigma_a(x, x')$ and ordinary differentiation with the covariant derivative. That is, given a bitensor $T_{\underline{a}}(x, x')$, where $\underline{a} = a_1 \cdots a_m$ is a list of unprimed indices, we perform the expansion

$$T_{\underline{a}}(x, x') = \sum_{k=0}^{\infty} \frac{(-1)^k}{k!} t_{\underline{a}b_1 \cdots b_k}(x) \sigma^{b_1} \cdots \sigma^{b_k}. \tag{1.11}$$

We can then solve for the expansion coefficients by repeated covariant differentiation and taking the coinciding point limit. Namely, it follows from (1.11) that

$$t_{\underline{a}} = [T_{\underline{a}}], \quad t_{\underline{a}b_1} = [T_{\underline{a};b_1}] - t_{\underline{a};b_1}, \quad t_{\underline{a}b_1 b_2} = [T_{\underline{a};b_1 b_2}] - t_{\underline{a};b_1 b_2} - t_{\underline{a}b_1;b_2} - t_{\underline{a}b_2;b_1},$$

etc. If T is a bitensor that also has primed indices, one first has to transport the primed indices to unprimed ones using the parallel propagator $g^a{}_{b'}$.

   Of course, just like any other Taylor expansion, the covariant expansion of a smooth bitensor does in general not converge and if it converges it is not guaranteed to converge to the bitensor that is being expanded. Only if the bitensor and the

metric are analytic, we will not encounter these difficulties. However, even then the covariant expansion can be a very useful *asymptotic* expansion and we usually truncate it after a finite number of terms. Therefore, in following we will work with the covariant series in a formal way and whenever we write an infinite sum we implicitly mean that the series is to be truncated and a finite remainder term is to be added.

1.4.4 Semi-recursive Avramidi method

As described, *e.g.*, in [69, 169], the 'naïve' recursive approach to calculate the expansion coefficients of a covariant expansion, as briefly sketched above, is inefficient and does not scale well to higher orders in the expansion because the calculation of coincidence limits becomes computationally prohibitive. An alternative, non-recursive and elegant method for the calculation of these coefficient was proposed by Avramidi [21, 22]. From the computational perspective, however, also this approach is not necessarily optimal as it does not always make good use of intermediate result leading to an algorithm which is space but not time efficient. A middle way, that we will present here, was implemented in [169] using a 'semi-recursive' method.

Avramidi's approach rests on the power series solution approach to solving differential equations. Therefore it can only be applied where the bitensor solves a differential equation, the *transport equations*.

With the *transport operators*

$$\nabla_\sigma \doteq \sigma^a \nabla_a \quad \text{and} \quad \nabla'_\sigma \doteq \sigma^{a'} \nabla_{a'}. \tag{1.12}$$

the world function (1.6) can be expressed as $(\nabla_\sigma - 2)\sigma = 0$ or $(\nabla'_\sigma - 2)\sigma = 0$. One can construct additional transport equations by differentiating these equations and commuting covariant derivatives at the cost of introducing curvature tensors. In particular we find [169]

$$\nabla'_\sigma \xi^{a'}{}_{b'} = \xi^{a'}{}_{b'} - \xi^{a'}{}_{c'} \xi^{c'}{}_{b'} - R^{a'}{}_{c'b'd'} \sigma^{c'} \sigma^{d'}, \tag{1.13a}$$

$$\nabla'_\sigma \Delta^{1/2} = \tfrac{1}{2} \Delta^{1/2} \left(n - \xi^{a'}{}_{a'} \right), \tag{1.13b}$$

where we have defined $\xi^{a'}{}_{b'} \doteq \sigma^{a'}{}_{b'}$ to avoid confusion later on.

In the next step we take the formal covariant expansion of the bitensor and use the transport equation to find relations between the expansion coefficients. Let us again take a bitensor $T_{\underline{a}}(x, x')$ with the expansion (1.11). Applying to it the transport operator ∇'_σ, we obtain formally

$$\nabla'_\sigma T_{\underline{a}}(x, x') = \sum_{m=0}^\infty \frac{(-1)^m}{m!} \nabla'_\sigma t_{\underline{a} b_1 \cdots b_m}(x) \sigma^{b_1} \cdots \sigma^{b_m}$$
$$= \sum_{m=1}^\infty \frac{(-1)^m}{m!} m\, t_{\underline{a} b_1 \cdots b_m}(x) \sigma^{b_1} \cdots \sigma^{b_m},$$

i.e., applying ∇'_σ to the m-th term is equivalent to multiplying it by m:

$$(\nabla'_\sigma T)_{(m)} = k T_{(m)}.$$

Contracting two bitensors $S_{\underline{a}}(x,x')$ and $T_{\underline{b}}(x,x')$, we further find

$$U_{\underline{ab}} = S_{\underline{ac}}T^{c}{}_{\underline{b}} = \sum_{m=0}^{\infty}\frac{(-1)^m}{m!}\sum_{k=0}^{k}\binom{m}{k}s_{\underline{a}cd_1\cdots d_k}t^{c}{}_{\underline{b}d_{k+1}\cdots d_m}\sigma^{d_1}\cdots\sigma^{d_m},$$

i.e., the m-th expansion coefficient of U is obtained from the lower order coefficients of S and T:

$$U_{\underline{ab}(m)} = \sum_{k=0}^{m}\binom{m}{k}S_{\underline{ac}(k)}T^{c}{}_{\underline{b}(m-k)}.$$

Moreover, we follow Avramidi and introduce

$$\mathcal{K}^{a}{}_{b(m)} \doteq R^{a}{}_{(c_1|b|c_2;c_3\cdots c_m)}\sigma^{c_1}\cdots\sigma^{c_m}$$

so that we can write the covariant expansion of $R^{a'}{}_{c'b'd'}\sigma^{c'}\sigma^{d'}$ as

$$g^{a}{}_{a'}g_{b}{}^{b'}R^{a'}{}_{c'b'd'}\sigma^{c'}\sigma^{d'} = \sum_{m=2}^{\infty}\frac{(-1)^m}{(m-2)!}\mathcal{K}^{a}{}_{b(m)}.$$

These formal manipulations and definitions can now be applied to find the covariant expansion coefficients of a bitensor. Here we will apply the method to $\xi^{a'}{}_{b'}$ and $\Delta^{1/2}$ with their transport equations (1.13). It follows from (1.7) and (1.8) that $\xi^{a}{}_{b(0)} = \delta^{a}_{b}$ and $\xi^{a}{}_{b(1)} = 0$. Then, for $m \geq 2$, we can easily find the relation

$$-(m+1)\xi^{a}{}_{b(0)} = \sum_{k=2}^{m-2}\binom{m}{k}\xi^{a}{}_{c(k)}\xi^{c}{}_{b(m-k)} + m(m-1)\mathcal{K}_{(m)}.$$

For the coefficients of the square-root of the van Vleck–Morette determinant one gets $\Delta^{1/2}_{(0)} = 1$ from (1.7) and (1.10) and

$$\Delta^{1/2}_{(m)} = -\frac{1}{2m}\sum_{k=0}^{m-2}\binom{m}{k}\Delta^{1/2}_{(k)}\xi^{a}{}_{a(m-k)}$$

for $m > 0$. Equivalent relations were found in [169] and implemented in the Mathematica™ package *CovariantSeries*.

1.4.5 Coordinate expansion of the world function

Our aim in this section is to find an expansion of Synge's world function σ in a coordinate neighbourhood. There are different ways to achieve this (see for example [83] for an approach using Riemannian normal coordinates). Here we use a (formal) power series Ansatz and write

$$\sigma(x,x') = \sum_{m=0}^{\infty}\frac{1}{m!}\varsigma_{\mu_1\cdots\mu_m}(x)\delta x^{\mu_1}\cdots\delta x^{\mu_m},$$

where we denote by $\delta x^{\mu} = (x'-x)^{\mu}$ the coordinate separation of the points x,x' in a chart.

The transport equation (1.6) can then be applied to obtain relations between the coefficients $\varsigma_{\mu_1\cdots\mu_m} = \varsigma_{(\mu_1\cdots\mu_m)}$. As consequence of (1.7) and $[\sigma_{\mu}] = 0$, the first three

coefficients are

$$\varsigma = 0, \quad \varsigma_\mu = 0 \quad \text{and} \quad \varsigma_{\mu\nu} = g_{\mu\nu}.$$

This can be used to derive the relation

$$2(1-m)\varsigma_{\mu_1\cdots\mu_m} = \sum_{k=2}^{m-2}\binom{m}{k}g^{\nu\rho}(\varsigma_{(\mu_1\cdots\mu_k|,\nu} - \varsigma_{(\mu_1\cdots\mu_k|\nu})(\varsigma_{\mu_{k+1}\cdots\mu_m),\rho} - \varsigma_{\mu_{k+1}\cdots\mu_m)\rho})$$
$$- 2m\varsigma_{(\mu_1\cdots\mu_{m-1},\mu_m)} \tag{1.14}$$

for $m > 2$ after a cumbersome but straightforward calculation.

It is not difficult to implement (1.14) efficiently in a modern computer algebra system; it involves multiplication, transposition and symmetrization of multidimensional arrays and partial derivatives on the components of these arrays. Making use of the symmetry of the coefficients, can reduce computation time and memory usage for higher order coefficients, especially for 'complicated' metrics, significantly. Even if these symmetries are not used, this method appears to be more efficient than the method of [83], as it does not require to compute Riemann normal coordinates first. In fact, because the Riemann normal coordinates are given by the derivative of the world function, the coefficients $\varsigma_{\mu_1\cdots\mu_m}$ can also be used to compute the expansion of Riemann normal coordinates:

$$\sigma_\mu(x,x') = g_{\mu\nu}\delta x^\nu + \tfrac{1}{2}(\varsigma_{\nu\rho,\mu} + \varsigma_{\mu\nu\rho})\delta x^\nu\delta x^\rho + \tfrac{1}{3!}(\varsigma_{\nu\rho\lambda,\mu} + \varsigma_{\mu\nu\rho\lambda})\delta x^\nu\delta x^\rho\delta x^\lambda + \cdots$$

Again, this method appears more direct and faster than the one described in [40].

For example, for the Friedmann-Lemaître-Robertson-Walker metric in cosmological time (see Sect. 2.3.1 for more details),

$$g = -\mathrm{d}t \otimes \mathrm{d}t + a(t)^2\delta_{ij}\,\mathrm{d}x^i \otimes \mathrm{d}x^j,$$

one finds with $H = \dot{a}/a$

$$\begin{aligned}
2\sigma(x,x') = {}& -\delta t^2 + a^2\delta\vec{x}^2\big(1 + H\delta t + \tfrac{1}{3}(H^2+\dot{H})\delta t^2 + \tfrac{1}{12}a^2H^2\delta\vec{x}^2 \\
&+ \tfrac{1}{12}(2H\dot{H} + \ddot{H})\delta t^3 + \tfrac{1}{12}a^2H(2H^2 + \dot{H})\delta t\,\delta\vec{x}^2 \\
&+ \tfrac{1}{180}(-4H^4 - 8H^2\dot{H} + 2\dot{H}^2 + 6H\ddot{H} + 3\dddot{H})\delta t^4 \\
&+ \tfrac{1}{360}a^2(48H^4 + 74H^2\dot{H} + 8\dot{H}^2 + 9H\ddot{H})\delta t^2\delta\vec{x}^2 \\
&+ \tfrac{1}{360}a^4H^2(4H^2 + 3\dot{H}))\delta\vec{x}^4 + O(\delta x^7)
\end{aligned}$$

in agreement with [83]. Above we denoted by a dot derivatives with respect to cosmological time t and defined the coordinate separation $\delta t = (x'-x)^0$ and $\delta\vec{x}^2 = \delta_{ij}(x'-x)^i(x'-x)^j$. Transforming to conformal time, so that

$$g = a(\tau)^2\big(-\mathrm{d}\tau \otimes \mathrm{d}\tau + \delta_{ij}\,\mathrm{d}x^i \otimes \mathrm{d}x^j\big),$$

one finds with $\mathcal{H} = a'/a$

$$
\begin{aligned}
2a^{-2}\sigma(x,x') = {} & -\delta\tau^2 + \delta\vec{x}^2 - \mathcal{H}\delta\tau^3 + \mathcal{H}\delta\tau\delta\vec{x}^2 - \tfrac{1}{12}(7\mathcal{H}^2 + 4\mathcal{H}')\delta\tau^4 \\
& + \tfrac{1}{6}(3\mathcal{H}^2 + 2\mathcal{H}')\delta\tau^2\delta\vec{x}^2 + \tfrac{1}{12}\mathcal{H}^2\delta\vec{x}^4 - \tfrac{1}{12}(3\mathcal{H}^3 + 5\mathcal{H}\mathcal{H}' + \mathcal{H}'')\delta\tau^5 \\
& + \tfrac{1}{12}(2\mathcal{H}^3 + 4\mathcal{H}\mathcal{H}' + \mathcal{H}'')\delta\tau^3\delta\vec{x}^2 + \tfrac{1}{12}\mathcal{H}(\mathcal{H}^2 + \mathcal{H}')\delta\tau\delta\vec{x}^4 \\
& - \tfrac{1}{360}(31\mathcal{H}^4 + 101\mathcal{H}^2\mathcal{H}' + 28\mathcal{H}'^2 + 39\mathcal{H}\mathcal{H}'' + 6\mathcal{H}''')\delta\tau^6 \\
& + \tfrac{1}{360}(15\mathcal{H}^4 + 61\mathcal{H}^2\mathcal{H}' + 20\mathcal{H}'^2 + 30\mathcal{H}\mathcal{H}'' + 6\mathcal{H}''')\delta\tau^4\delta\vec{x}^2 + \\
& + \tfrac{1}{360}(15\mathcal{H}^4 + 37\mathcal{H}^2\mathcal{H}' + 8\mathcal{H}'^2 + 9\mathcal{H}\mathcal{H}'')\delta\tau^2\delta\vec{x}^4 \\
& + \tfrac{1}{360}(\mathcal{H}^4 + 3\mathcal{H}^2\mathcal{H}')\delta\vec{x}^6 + O(\delta x^7),
\end{aligned}
$$

where we denoted by a prime derivatives with respect to conformal time τ and defined $\delta\tau = (x' - x)^0$ and $\delta\vec{x}$ as above.

The method described above is sufficiently fast to be applied to more complicated spacetimes such as Schwarzschild or Kerr spacetimes. Nevertheless, for such spacetimes the expansions become so long, that they easily fill more than a page if printed up to sixth order. Therefore we will omit them here and suggest that the interested reader implements (1.14) in a computer algebra system.

2

Lorentzian geometry

Summary

Around the turn of the 20th century it became clear that several physical phenomena could not be satisfactorily described in the framework of Newtonian physics. After insights by Lorentz, Poincaré and others, a solution to some of these problems was found and published by Einstein in 1905 and coined special relativity. In its modern form the changes introduced by Einstein can be seen to arise from to the union of space and time in a single spacetime. In the absence of gravity or in the weak gravitational limit our universe appears to be well described by Minkowski spacetime, its causal structure and its symmetries.

The changes introduced by Minkwoski spacetime are generalized by the concept of Lorentzian spacetimes. This class of pseudo-Riemannian manifolds forms the basis of general relativity, introduced by Einstein in 1916, which takes the lessons learned from special relativity to formulate a geometric theory of gravitation. In Einstein's theory of gravitation the geometrical background of the Universe is itself dynamical and accounts for gravity via Einstein's equation. The total lack of static background with respect to which length and time measurements can be made, makes general relativity significantly more difficult than Newton's classical theory of gravitation.

Nevertheless, general relativity has made many predictions like gravitational redshift, gravitational lensing and, in a sense, the Big Bang, that have been confirmed with modern observations. It also forms the basis of modern cosmology which attempts to describe the evolution of the Universe on large scales from the time of the Big Bang until present times and predict its 'fate'.

This chapter will give a short introduction to the causal structure of Lorentzian spacetimes, to general relativity and to cosmology in three separate sections. The aim of the first section (Sect. 2.1) is the definition of globally hyperbolic spacetimes and Cauchy surfaces. Therefore it will define notions such as spacetimes, time functions, the causal past and future of a set, causality conditions, and the splitting of a spacetime into its temporal and spatial part. In the second section (Sect. 2.2) we give an overview of general relativity including the stress-energy tensor, the classical energy conditions, Einstein's equation and its fluid formulation, and the special case of de Sitter spacetime. Topics related to theoretical cosmology will be addressed in the third section (Sect. 2.3), where we discuss Friedmann–Lemaître–Robertson–Walker spacetimes and perturbations around it.

As we are presenting standard material, we will again mostly refrain from giving proofs or even sketching them. Excellent references for this chapter are the books [27, 80, 167, 216]: While [167] and in particular [27] contain detailed discussions of causal structure of Lorentzian spacetimes, [216] gives an good overall account of general relativity. The recent book [80] is, as the title already suggests, mostly focused on cosmology, and may be taken as a reference for the last section.

2.1 Causality

2.1.1 Causal structure

The metric g of an arbitrary Lorentzian manifold (M, g) distinguishes three regions in the tangent space: a nonzero $v \in T_x M$ is

$$
\begin{aligned}
\textit{spacelike} \quad &\text{if} \quad g_x(v, v) > 0, \\
\textit{lightlike} \text{ or } \textit{null} \quad &\text{if} \quad g_x(v, v) = 0, \\
\textit{timelike} \quad &\text{if} \quad g_x(v, v) < 0, \\
\textit{causal} \quad &\text{if} \quad g_x(v, v) \leq 0.
\end{aligned}
$$

Furthermore, we define the zero vector $v = 0$ as spacelike, as *e.g.* in [167].

Given a *time-orientation u* on M, *viz.*, a smooth unit timelike vector field such that $g(u, u) = -1$, a causal vector $v \in T_x M$ is

$$
\begin{aligned}
\textit{future-directed} \quad &\text{if} \quad g_x\big(u(x), v\big) < 0, \\
\textit{past-directed} \quad &\text{if} \quad g_x\big(u(x), v\big) > 0.
\end{aligned}
$$

One often chooses the time-orientation to be the velocity of a physical fluid. Not every Lorentzian manifold can be given a time-orientation and we say that a Lorentzian manifold is *time-orientable* is such a vector field exists.

All these notions extend naturally to the cotangent bundle T^*M, sections of both TM and T^*M and curves. That is, as seen in Fig. 2.1, to each point $x \in M$ we can draw a double cone inside M whose surface is generated by the lightlike curves passing through x; points lying inside the double cone are lightlike to x, while points lying outside are spacelike to x.

A continuous function $t : M \to \mathbb{R}$ that is strictly increasing along every future-directed causal curve is called a *time function*. Clearly, if t is a time function, then its gradient yields a time-orientation which is orthogonal to the level surfaces

$$
t^{-1}(c) = \{x \in M \mid f(x) = c\}, \quad c \in \mathbb{R}.
$$

These level surfaces give a foliation of the manifold such that each leaf is spacelike and intersected at most once by every causal curve.

The *chronological future $I^+(U)$* (*chronological past $I^-(U)$*) of a subset $U \subset M$ is defined as the set of points which can be reached from U by future-directed (past-directed) timelike curves. Similarly, the *causal future $J^+(U)$* (*causal past $J^-(U)$*) of a subset $U \subset M$ is defined as the set of points which can be reached from U by future-directed (past-directed) causal curves; their union $J(U) \doteq J^+(U) \cup J^-(U)$ is the *causal shadow* of U. If the set U consists of only one point $U = \{x\}$, we write $I^\pm(x)$ and $J^\pm(x)$. Note that $I^\pm(x)$ is always open, whereas $J^\pm(x)$ is not necessarily closed.[1] We say that two subsets U and V are *causally separated* (in symbols: $U \times V$) if $U \cap J(V) = \emptyset$. These definitions are illustrated in Fig. 2.2.

A *world line* or *observer* is a timelike future-directed curve $\tau \mapsto \gamma(\tau)$ such that $g(\dot\gamma, \dot\gamma) = -1$ and the curve parameter τ is called the *proper time* of γ. The observer γ

---

[1] In globally hyperbolic spacetimes (to be defined below) $J^\pm(x)$ is always closed.
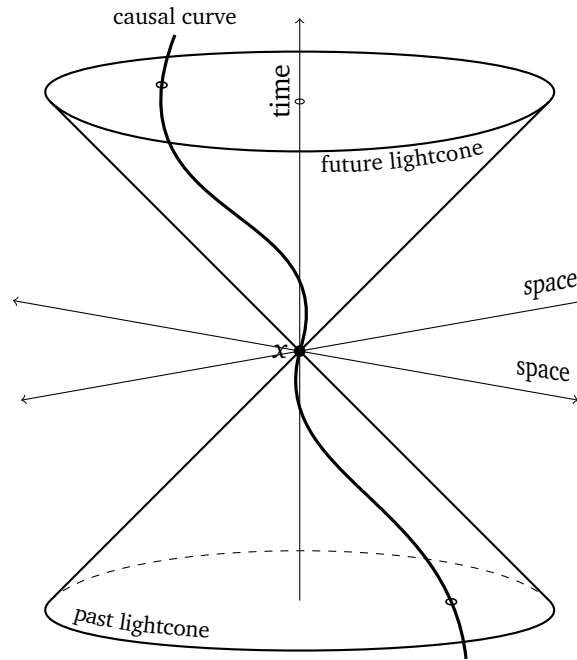
Figure 2.1. Lightcone of a point in a three-dimensional Minkowski spacetime.

is said to be freely falling, *i.e.*, moving only under the influence of gravity, if γ is a geodesic.

The causal structure of a Lorentzian manifold is physically significant because it is used to imply that two *events*, *i.e.*, two points on the manifold, can only influence each other if one is in the lightcone of the other. In other words, the Lorentzian causal structure encodes the finiteness of signalling speeds. This is distinctively different from Newtonian physics, where one event at an absolute time t_0 influences all other events at a later time $t > t_0$, no matter the spatial separation.

2.1.2 Covariant splitting

Let (M, g) be a $(1 + n)$-dimensional Lorentzian manifold with a time-orientation u. Observe that the integral curves of u can be understood to define a preferred direction of motion. It is clear that the integral curves of u can be parametrized such that they are world lines. A local frame such that the components of u are

$$u^\mu = (1, 0, \dots, 0)$$

is called a *comoving frame*.

Orthogonal to u at each point are the *rest spaces* of the associated observer. An induced Riemannian metric tensor for these n-spaces is given by the projected tensor

$$h \doteq g + u^\flat \otimes u^\flat,$$

which has the following properties: $h_{ab} u^b = 0$, $h_a{}^c h_c{}^b = h_a{}^b$, $h_a{}^a = n$.

By means of the projection h and the time-orientation u, we can decompose any tensor into its temporal and spatial parts with respect to the observer. In particular,
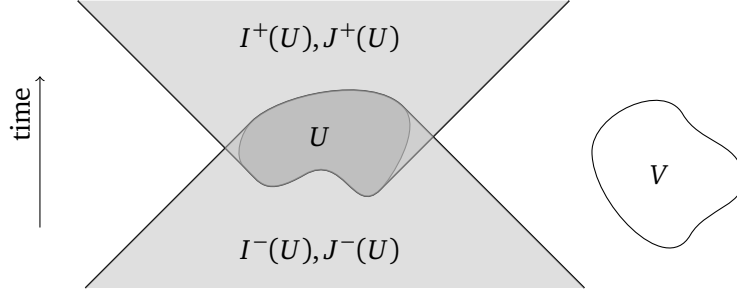
Figure 2.2. Causal past and future of a set and a causally separated set.

we can define the *temporal* and *spatial covariant derivatives* of any tensor $S^{a\cdots}{}_{b\cdots}$ by

$$\dot{S}^{a\cdots}{}_{b\cdots} \doteq u^c \nabla_c S^{a\cdots}{}_{b\cdots},$$

$$\overline{\nabla}_c S^{a\cdots}{}_{b\cdots} \doteq h^a{}_p \cdots h_b{}^q \cdots h_c{}^r \nabla_r S^{p\cdots}{}_{q\cdots}.$$

Moreover, define the *projected symmetric trace-free (PSTF) parts* of a vector field v^a and a tensor S_{ab} by

$$v^{\langle a \rangle} \doteq h^a{}_b v^b, \quad S_{\langle ab \rangle} \doteq \left(h_{(a}{}^c h_{b)}{}^d - \tfrac{1}{n} h_{ab} h^{cd} \right) S_{cd}.$$

The kinematics of the world lines given by u is characterized by $\nabla_b u_a$ which can be decomposed as

$$\nabla_b u_a = \overline{\nabla}_b u_a - \dot{u}_a u_b = \omega_{ab} + \Theta_{ab} - \dot{u}_a u_b = \omega_{ab} + \sigma_{ab} + \tfrac{1}{n}\Theta h_{ab} - \dot{u}_a u_b, \quad (2.1)$$

with

$$\omega_{ab} \doteq \overline{\nabla}_{[b} u_{a]}, \quad \sigma_{ab} \doteq \overline{\nabla}_{(a} u_{b)}, \quad \Theta \doteq \Theta^a{}_a = \nabla_a u^a, \quad \Theta_{ab} \doteq \overline{\nabla}_{\langle a} u_{b \rangle},$$

where ω_{ab} denotes the *vorticity (twist) tensor*, Θ_{ab} the *expansion tensor*, Θ the *(volume) expansion scalar*, σ_{ab} the *shear tensor*, and $\dot{u}_a$ the *acceleration*. The expansion scalar measures the separation of neighbouring observers and can be used to introduce a length scale a via the definition

$$\frac{\dot{a}}{a} \doteq \frac{\Theta}{n}.$$

Furthermore, we shall define the magnitudes ω and σ of the vorticity and the shear tensor by

$$\omega^2 \doteq \tfrac{1}{2}\omega_{ab}\omega^{ab}, \quad \sigma^2 \doteq \tfrac{1}{2}\sigma_{ab}\sigma^{ab}.$$

Using the definition of the Ricci tensor, we find $2\nabla_{[a}\nabla_{b]}u^a u^b = R_{ab}u^a u^b$ which, together with (2.1), yields an equation describing the dynamics of world line

$$-R_{ab}u^a u^b = \dot{\Theta} + \tfrac{1}{n}\Theta^2 + 2(\sigma^2 - \omega^2) - \dot{u}^a{}_{;a}, \quad (2.2)$$

where $R_{ab}u^a u^b$ is sometimes called the *Raychaudhuri scalar*.

If the vorticity tensor vanishes, then, by Frobenius' theorem, u will be orthogonal to a foliation of the spacetime by n-dimensional Euclidean hypersurfaces Σ whose metric is given by (the pullback of) h. If the manifold is also simply connected so

(a)   (b)
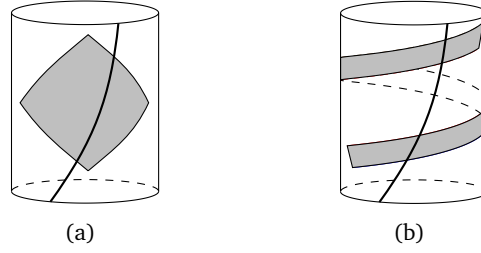
Figure 2.3. Sets on the two-dimensional Einstein cylinder: (a) is causally convex, while (b) is not, as illustrated by the same causal curve in both figures.

that its first de Rham cohomology group vanishes, then there exists even a time function.[2] Furthermore, in the case of vanishing vorticity, $\overline{\nabla}$ can be directly identified with the Levi-Civita connection on the spatial slices Σ with metric tensor h. Therefore it can be used to derive relations between the Riemann curvature $\overline{R}_{abcd}$ of (Σ, h) and the Riemann curvature R_{abcd} of the whole manifold (M, g). The *Gauss–Codacci equation* [216, Chap. 10] is

$$\overline{R}_{abcd} = h_a{}^p h_b{}^q h_c{}^r h_d{}^s R_{pqrs} - 2\Theta_{a[c}\Theta_{d]b}, \tag{2.3}$$

where we have expressed the extrinsic curvature via by the expansion tensor. Upon contraction with the projection h the spatial Riemann tensor leads to the projected Ricci tensor and Ricci scalar.

2.1.3   Cauchy surfaces

Although many results hold in greater generality, for physical reasons we will from now restrict our attention to a subclass of four-dimensional Lorentzian manifolds:

Definition 2.1. *A* spacetime *is a connected, oriented (positively or negatively oriented with respect to the volume form induced by the metric) and time-oriented, four-dimensional smooth[3] Lorentzian manifold $(M, g, \pm, u)$. Usually we omit orientation and time-orientation and identify a spacetime with the underlying Lorentzian manifold (M, g); sometimes, when no confusion can arise, we will even drop the metric and say that M is a spacetime.*

Not all spacetimes are of physical significance because they admit features like closed timelike curves (*i.e.*, 'time machines') that are usually considered unphysical. A spacetime that does not have any closed timelike curves (*i.e.*, $x \notin I^+(x)$ for all $x \in M$) is said to satisfy the *chronological condition*; no compact spacetime satisfies the chronological condition. A slightly stronger notion is the *causality condition*, which forbids the existence of closed *causal* curves.

Given an open set $U \subset M$, it is called *causally convex* if the intersection of any causal curve with U is a connected set (possibly the empty set). That is, a causal

---

[2]A local time function will always exist if the vorticity is vanishing because every manifold is locally contractible.

[3]On some occasions we will work with spacetimes whose metric is not smooth. In all these cases we will mention the regularity of the metric explicitly.

curve intersecting U cannot leave the set and enter it again (Fig. 2.3). If for every neighbourhood V of $x \in M$ there is a causally convex neighbourhood $U \subset V$ of x, then the *strong causality condition* holds at x. A spacetime (M, g) is called *strongly causal* if it is strongly causal at every point. Finally, if (M, g) is strongly causal and $J^+(x) \cap J^-(y)$ is compact for every pair $x, y \in M$, it is called *globally hyperbolic*.

A *Cauchy surface* is a subset $\Sigma \subset M$ which is intersected exactly once by every inextendible timelike curve (and at least once by every inextendible causal curve). Therefore, the causal shadow of a Cauchy surface is the entire spacetime. Note that a Cauchy surface may be non-spacelike and non-smooth. Our ability to pose a Cauchy problem on a spacetime requires the existence of Cauchy surfaces.

The existence of a Cauchy surface imposes strong conditions on the causality of the spacetime. Given a spacetime, the following statements are equivalent [35, 167]: the spacetime

  (a) is globally hyperbolic,

  (b) admits a Cauchy surface,

  (c) admits a smooth time function compatible with the time-orientation.

Therefore we will be mostly interested in globally hyperbolic spacetimes in the following. Every globally hyperbolic spacetime (M, g) is diffeomorphic to $\mathbb{R} \times \Sigma$, where Σ is diffeomorphic to a smooth spacelike Cauchy surface of M [34]. If we denote by t the time function on M, then the level sets of t are isometric to (Σ, g_t), where g_t is a Riemannian metric on Σ depending smoothly on t. In fact, (M, g) is isometric to $\mathbb{R} \times \Sigma$ with the metric

$$-\beta \, \mathrm{d}t \otimes \mathrm{d}t + g_t,$$

where β is a smooth function on M.

2.2   General relativity

Throughout this section let $(M, g, \pm, u)$ be an arbitrary spacetime, unless otherwise specified.

2.2.1   The stress-energy tensor

General relativity describes the interaction of classical matter with the geometrical structure of the Universe. Giving a precise definition of matter is difficult if not impossible. Mathematically matter is described in general relativity as a covariant or contravariant symmetric 2-tensor field T_{ab} or T^{ab}, the *energy-momentum* or *stress-energy tensor* which is covariantly conserved

$$\nabla^b T_{ab} = 0.$$

The physical content of the stress-energy tensor T_{ab} becomes clearer once we perform a covariant splitting relative to u and decompose T_{ab} as

$$T_{ab} = \rho u_a u_b + p h_{ab} + 2 q_{(a} u_{b)} + \pi_{ab}, \tag{2.4}$$

with

$$\rho \doteq T_{ab}u^a u^b, \quad p \doteq \tfrac{1}{3}T_{ab}h^{ab}, \quad q_a \doteq -T_{\langle a \rangle b}u^b, \quad \pi_{ab} \doteq T_{\langle ab \rangle}, \qquad (2.5)$$

where ρ denotes the *energy density*, p the *pressure*, q_a the *momentum density* (*viz.*, the dissipation relative to u^a) and π_{ab} the *anisotropic stress*. By definition of these quantities, the *trace of the stress-energy tensor* is given by

$$T = T^a{}_a = -\rho + 3p.$$

Instead of working with the stress-energy tensor it is thus possible to work with these four quantities and the equations of state relating them. For example, in this decomposition the conservation equation $T^{ab}{}_{;b} = 0$ splits into the *energy conservation* and the *momentum conservation equation*

$$\dot{\rho} + (\rho + p)\Theta + \pi^{ab}\sigma_{ab} + \overline{\nabla}_a q^a + 2\dot{u}_a q^a = 0, \qquad (2.6a)$$

$$\overline{\nabla}_a p + (\rho + p)\dot{u}_a + \pi_{ab}\dot{u}^b + \overline{\nabla}^b \pi_{ab} + \dot{q}_a + \tfrac{4}{3}\Theta q_a + (\sigma_{ab} + \omega_{ab})q^b = 0, \qquad (2.6b)$$

which are the familiar equations for an inertial observer on Minkowski spacetime.

If the anisotropic terms in (2.4) vanish ($q_a = 0 = \pi_{ab}$), the stress-energy takes the *perfect fluid* form

$$T_{ab} = \rho u_a u_b + p h_{ab} = (\rho + p)u_a u_b + p g_{ab} \qquad (2.7)$$

and the conservation equations (2.6) reduce to

$$\dot{\rho} + (\rho + p)\Theta = 0, \qquad (2.8a)$$

$$\overline{\nabla}_a p + (\rho + p)\dot{u}_a = 0. \qquad (2.8b)$$

One further distinguishes between different forms of the general equation of state $p = p(\rho, s)$, where s is the medium's specific entropy. Namely, if $p = p(\rho)$, one speaks of a *barotropic fluid*, and if $p = 0$, we have *pressure-free matter*, also called *'dust'*.

While general relativity imposes no *a priori* constraints on the form of the matter content, many possibilities can be considered unphysical in classical physics. The most common *energy conditions* are:

(NEC) *null energy condition*

$$T_{ab}v^a v^b \geq 0 \qquad \text{for all lightlike } v^a,$$

i.e., no negative energy densities along any lightray;

(WEC) *weak energy condition*

$$T_{ab}v^a v^b \geq 0 \qquad \text{for all timelike } v^a,$$

i.e., no observer detects negative energy densities;

(DEC) *dominant energy condition*

$$T_{ab}v^a w^b \geq 0 \qquad \text{for all future-pointing timelike } v^a, w^b,$$

i.e., the stress-energy flux is causal;

(SEC) *strong energy condition*

$$T_{ab}v^a v^b - \tfrac{1}{2}T \geq 0 \qquad \text{for all unit timelike } v^a,$$

i.e., gravity is attractive if $\Lambda = 0$, see (2.13).

By continuity, WEC implies NEC and it is also not difficult to see that DEC implies WEC. Moreover, the SEC *does not* imply the WEC but only the NEC. The reverse implications are generally not true. Note that the strong energy condition is too strong for many physically relevant scenarios.

A generic feature of quantum field theory is that none of the energy conditions above will hold, even in an averaged sense, because of the Reeh–Schlieder theorem [110, 192]. Instead one finds lower bounds on the averaged energy density, called *quantum energy inequalities*, see [87, 88] for a review of the subject.

2.2.2   Einstein's equation

Einstein's equation with a cosmological constant are

$$R_{ab} - \frac{1}{2}R g_{ab} + \Lambda g_{ab} \doteq G_{ab} + \Lambda g_{ab} = \frac{8\pi G}{c^4}T_{ab}, \qquad (2.9)$$

where G_{ab} is called the *Einstein (curvature) tensor* and Λ the *cosmological constant*. The constants on the right-hand side are *Newton's gravitational constant* G and the *speed of light c*; we shall always choose units such that $8\pi G = c = 1$. Often we will also work with the trace of (2.9):

$$-R + 4\Lambda = T. \qquad (2.10)$$

Sometimes one absorbs the cosmological constant into the stress-energy tensor to emphasize its non-geometric nature.

Note that Einstein's tensor is covariantly conserved and symmetric, *i.e.*,

$$\nabla^b G_{ab} = 0 \quad \text{and} \quad G_{ab} = G_{(ab)}, \qquad (2.11)$$

so that the left-hand side of (2.9) is consistent with the right-hand side and gives a second-order differential equation in the metric. In fact, we can derive (2.9) from the assumption that the stress-energy tensor of a matter field should be the source of a gravitational potential (the metric tensor) in a second-order differential equation. Since the stress-energy tensor is conserved and symmetric, (2.9) is the only possibility.

Combining (2.9) with (2.10), Einstein's equations can be recast into the equivalent form $R_{ab} - \Lambda g_{ab} = T_{ab} - \tfrac{1}{2}T g_{ab}$. Together with the imperfect fluid form (2.4) of the stress-energy tensor this equation yields

$$R_{ab} - \Lambda g_{ab} = \tfrac{1}{2}(\rho + 3p)u_a u_b + \tfrac{1}{2}(\rho - p)h_{ab} + 2q_{(a}u_{b)} + \pi_{ab}.$$

Contractions with the time-orientation u^a and the associated projector h_{ab} to the orthogonal surfaces then give the three equation

$$R_{ab}u^a u^b = \tfrac{1}{2}(\rho + 3p) - \Lambda, \qquad (2.12a)$$

$$R_{bc}h_a{}^b u^c = -q_a, \qquad (2.12b)$$

$$R_{cd}h_a{}^c h_b{}^d = \tfrac{1}{2}(\rho - p)h_{ab} + \Lambda h_{ab} + \pi_{ab}. \qquad (2.12c)$$

The Raychaudhuri scalar (2.2) attains the physical meaning of the active grav-
itational energy by (2.12a) and thus we can state the *Raychaudhuri equation* as:

$$\dot{\Theta} + \tfrac{1}{3}\Theta^2 + 2(\sigma^2 - \omega^2) - \overline{\nabla}_a \dot{u}^a - \dot{u}_a \dot{u}^a + \tfrac{1}{2}(\rho + 3p) - \Lambda = 0. \qquad (2.13)$$

Hence we see that expansion, shear and matter (satisfying the strong energy con-
dition) promote gravitational collapse, whereas a positive cosmological constant,
vorticity and positive acceleration (due to non-gravitational forces inside the medium)
oppose gravitational collapse. One can also derive differential equations for the shear
and the vorticity [212].

2.2.3 De Sitter spacetime

A Lorentzian manifold (M, g) is called a *vacuum solutions* to Einstein's equation (2.9)
if both T_{ab} and Λ vanish globally. That is, such a solution satisfies

$$R_{ab} = 0.$$

The vacuum Einstein equation is the most studied special case of the Einstein equation
and many important and instructive examples fall into this class. The most basic
solution is Minkowski spacetime (M, η), see (1.2), which describes a featureless
empty universe. A rotation-symmetric vacuum solution is given by the famous
Schwarzschild solution.

If we also allow for a cosmological constant but keep $T_{ab} = 0$, we need to solve

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

The maximally symmetric solutions of this equation fall into three classes depending
on the sign of Λ. We focus here on the case $\Lambda > 0$ called *de Sitter spacetime*; for $\Lambda = 0$
one obtains Minkowski spacetime and for $\Lambda < 0$ the so-called *anti-de Sitter spacetime*.

Four-dimensional de Sitter spacetime is the hyperboloidal submanifold of five-
dimensional Minkowski spacetime with coordinates $(y^0, y^1, \ldots, y^4)$ that satisfies the
equation

$$-(y^0)^2 + (y^1)^2 + \cdots + (y^4)^2 = H^{-2}, \quad \Lambda = 3H^2, \qquad (2.14)$$

where $H > 0$ is called the *Hubble constant*; thus de Sitter space is topologically $\mathbb{R} \times S^3$.
The pullback of the Minkowski metric to this space yields a Lorentzian metric.

A global coordinate chart with coordinates $(t, \chi, \theta, \varphi)$ can be defined by

$$y^0 = H^{-1} \sinh(Ht), \quad y^i = H^{-1} \cosh(Ht) z^i,$$

where $z^i = z^i(\chi, \theta, \varphi)$, $i = 1, 2, 3, 4$, are the usual spherical coordinates on S^3 with
unit radius. In these coordinates the induced metric on de Sitter spacetime reads

$$\begin{aligned}
g &= -\mathrm{d}t \otimes \mathrm{d}t + H^{-2} \cosh(Ht)\big(\mathrm{d}\chi \otimes \mathrm{d}\chi + \sin^2\chi\,(\mathrm{d}\theta \otimes \mathrm{d}\theta + \sin^2\theta\,\mathrm{d}\varphi \otimes \mathrm{d}\varphi)\big) \\
&= -\mathrm{d}t \otimes \mathrm{d}t + H^{-2} \cosh(Ht)\, g_{\mathbb{S}^3},
\end{aligned}$$

where $g_{\mathbb{S}^3}$ denotes the standard metric on the 3-sphere. It is clear, that de Sitter
spacetime is globally hyperbolic and the constant time hypersurfaces are Cauchy
surfaces.

Another coordinate chart with coordinates (t, x^1, x^2, x^3) is given by

$$y^0 = H^{-1}\sinh(Ht) + \tfrac{1}{2}He^{Ht}r^2,$$
$$y^i = e^{Ht}x^i,$$
$$y^4 = H^{-1}\cosh(Ht) - \tfrac{1}{2}He^{Ht}r^2,$$

where $i = 1,2,3$ and $r^2 = (x^1)^2 + (x^2)^2 + (x^3)^2$. It covers only the half of de Sitter spacetime which satisfies $y^0 + y^4 \geq 0$ and is called the *cosmological patch* or *cosmological chart*; it is diffeomorphic to four-dimensional Euclidean space. Within the cosmological chart, the metric reads

$$g = -dt \otimes dt + e^{2Ht}\delta_{ij}\,dx^i \otimes dx^j = (H\tau)^{-2}(-d\tau \otimes d\tau + \delta_{ij}\,dx^i \otimes dx^j),$$

where we define the *conformal time* $\tau \in (0, \infty]$ via

$$t \mapsto \tau(t) = -\int_t^\infty a(t')^{-1}\,dt'.$$

Note that this metric is a special case of a Friedmann-Lemaître-Robertson-Walker metric, to be discussed in more generality in Sect. 2.3.

On de Sitter spacetime, Synge's world function is known in closed form and is closely related to the geodesic distance on five-dimensional Minkowski spacetime. In fact, since the chord length between two points x, x' on de Sitter space considered as the hyperboloid (2.14) is

$$Z(x, x') \doteq H^2\eta_{\alpha\beta}y^\alpha(x)y^\beta(x'), \tag{2.15}$$

Synge's world function on de Sitter spacetime is given by

$$\cos\left(H\sqrt{2\sigma(x, x')}\right) = Z(x, x') \tag{2.16}$$

for $|Z| \leq 1$, *i.e.*, for x' are not timelike to x. Equation (2.16) can be analytically continued to timelike separated points x, x', whence we find

$$\cosh\left(H\sqrt{-2\sigma(x, x')}\right) = Z(x, x') \quad \text{for } |Z| > 1.$$

The possible values for the chord length Z are illustrated in a conformal diagram of de Sitter spacetime in Fig. 2.4. In the cosmological chart, the function Z attains the simple form

$$Z(x, x') = 1 + \frac{(\tau - \tau')^2 - (\vec{x} - \vec{x}')^2}{2\tau\tau'} = \frac{\tau^2 + \tau'^2 - (\vec{x} - \vec{x}')^2}{2\tau\tau'}, \tag{2.17}$$

where $x = (\tau, \vec{x})$ and $x' = (\tau', \vec{x}')$. Note that the fraction on the right-hand side is a rescaling of Synge's world function on Minkowski spacetime by $-2\tau\tau'$.

2.3 Cosmology

When studying cosmological problems on usually describes the Universe by a homogeneous and isotropic spacetime, *i.e.*, one assumes the Friedmann-Lemaître-Robertson-
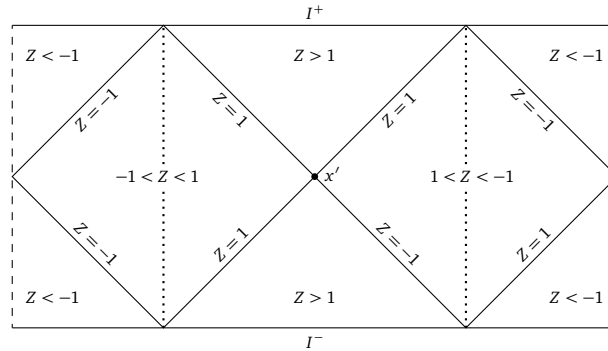
Figure 2.4. De Sitter spacetime conformally mapped to a cylinder. The cylinder is unwrapped and the left and right dotted edges must be identified. x' is any (fixed) point of de Sitter spacetime and $Z(x, x')$ is shown for all choices of x. The dotted line represents $Z = 0$. See also [10, Fig. 2].

Walker model. This standard prescription underpins the so called standard model of cosmology, the ΛCDM model of cold dark matter with a cosmological constant. However, from the presence of structure in the Universe (*e.g.*, galaxy clusters, galaxies, stars, *etc.*) we can directly deduce that the Universe is *neither* homogeneous *nor* isotropic. Instead it is believed that one can describe the Universe as nearly homogeneous and isotropic on cosmological scales so that observed Universe can be modelled as a perturbation around a FLRW spacetime and homogeneity and isotropy hold in an averaged sense.

The cosmic microwave background (CMB) and the galaxy distribution are often believed to give a direct justification of this idea, but since all observations are along the past light cone and do not measure an instantaneous spatial surface one can only directly observe isotropy. The link to homogeneity is less clear but if *all* observers measure an isotropic CMB, it can be shown that the spacetime is also homogeneous. In this respect, the current cosmological model is heavily influenced by the philosophical paradigm in cosmology, the Copernican principle, that we live in no distinguished region of the Universe and that other observers would observe the same. But even under this additional assumption it is not completely clear that near homogeneity follows since the CMB is not exactly but only nearly isotropic; see [78, 80, 148, 205] for a discussion of this issue.

Ignoring these shortcomings, we follow the standard approach and assume that we live in an 'almost'-FLRW spacetime, *i.e.*, a universe which is correctly described by a perturbation around a FLRW background. Because such an assignment of a background spacetime to the physical perturbed spacetime is not unique, one has to deal with a gauge problem.

Accordingly, we will begin by introducing the FLRW model and its properties. Then we will study the general gauge problem and its application in the case of a perturbation around a FLRW spacetime.

2.3.1 Friedmann-Lemaître-Robertson-Walker spacetimes

One of the simplest solutions of the Einstein equation is the *Friedmann-Lemaître-Robertson-Walker (FLRW) solution*, which describes a *homogeneous* and *isotropic*,

expanding or contracting universe. Let (M, g) be a spacetime, which is still to be determined, with a preferred flow (time-orientation) u. As a consequence of (spatial) isotropy, *i.e.*, the absence of any preferred (spatial) direction, vorticity, shear and acceleration have to vanish:

$$\omega_{ab} = 0, \quad \sigma_{ab} = 0, \quad \dot{u}_a = 0.$$

Also the anisotropic terms of the stress-energy tensor have to vanish, *i.e.*,

$$q_a = 0, \quad \pi_{ab} = 0$$

so that the stress-energy tensor takes the perfect fluid form (2.7). Consequently the FLRW model is completely determined by its energy density ρ, pressure p and the expansion Θ.

Since vorticity and acceleration vanish, the spacetime (M, g) is foliated by surfaces Σ orthogonal to u, which are required to be homogeneous by assumption and thus ρ, p and Θ are constant on these surfaces. For the same reason there exists locally a time function t, called *cosmological time*, that measures proper time, defined up to a constant shift, such that $u_a = -\nabla_a t$. Henceforth we will always assume the time function exists globally so that the resulting spacetime is stably causal and, in particular, globally hyperbolic.

One can show that the projected Ricci tensor, *i.e.*, the Ricci tensor for the spatial slices Σ (*cf.* (2.3)), simplifies significantly to

$$\overline{R}_{ab} = \tfrac{1}{3}\overline{R}h_{ab} = \tfrac{2}{3}(\rho + \Lambda - \tfrac{1}{3}\Theta^2)h_{ab}.$$

We may recast this equation into the more familiar form of the *first Friedmann equation*

$$H^2 = \tfrac{1}{3}\left(\rho + \Lambda - \tfrac{1}{2}\overline{R}\right). \tag{2.18}$$

with the famous *Hubble parameter* or *Hubble function* $H \doteq \Theta/3 = \dot{a}/a$. The *second Friedmann equation* is a special case of the Raychaudhuri equation (2.13) and reads

$$\dot{H} + H^2 + \tfrac{1}{6}(\rho + 3p) - \tfrac{1}{3}\Lambda = 0. \tag{2.19}$$

These two equations can be complemented with the energy conservation equation for the perfect fluid (2.8a) to show that $\overline{R}a^2$ is a constant.

We define $K \doteq \overline{R}a^2/3$ and notice that the initial value for the scale factor $a > 0$ is arbitrary so that we can restrict its value to $K = -1, 0, +1$. The sign of K determines the local geometry of the spatial sections: $K = -1, 0, +1$ correspond respectively to a hyperbolic, a flat and a elliptic geometry. The topology of the spatial sections is not completely determined by K and, in fact, there are many possibilities. While $K = +1$ implies that the spatial sections are compact, both compact and non-compact spatial section are possible for $K = -1, 0$.

Which of these three distinct values for K is realized depends on the energy density contained in the universe. If the energy density takes on the critical value $\rho = \rho_c \doteq 3H^2 - \Lambda$, the spatial surfaces will be flat, while $\rho > \rho_c$ leads to a spherical and $\rho < \rho_c$ leads to a hyperbolical geometry. Furthermore, according to (2.19), an

accelerating expansion ($\ddot{a} > 0$) occurs when $\rho + 3p < 0$ (assuming $\Lambda = 0$), *i.e.,* when the strong energy condition is violated.

A metric tensor realizing the FLRW universe in a local comoving coordinate frame $x^\mu = (t, r, \theta, \phi)$ with respect to u^a (*i.e.,* $u = \partial_t$) is given (locally) by the *FLRW metric*

$$g = -\mathrm{d}t \otimes \mathrm{d}t + a(t)^2 \big(\mathrm{d}r \otimes \mathrm{d}r + f_K(r)^2 (\mathrm{d}\theta \otimes \mathrm{d}\theta + \sin^2\theta\,\mathrm{d}\phi \otimes \mathrm{d}\phi)\big), \qquad (2.20)$$

where

$$f_K(r) \doteq \begin{cases} \sin r & \text{for } K = +1 \\ r & \text{for } K = 0 \\ \sinh r & \text{for } K = -1. \end{cases}$$

Therefore, a spacetime is a FLRW spacetime if and only if its metric attains locally the form (2.20) in some coordinate system and the time-orientation is given by $\mathrm{d}t$.

Let us define *conformal time* τ via $\mathrm{d}\tau = \mathrm{d}t/a$. That is we set

$$\tau(t) = \tau_0 - \int_{t_0}^{t} \frac{1}{a(t')}\,\mathrm{d}t' \quad \text{or} \quad \tau(t) = \int_{t}^{t_1} \frac{1}{a(t')}\,\mathrm{d}t' - \tau_1$$

with arbitrary τ_0, τ_1 and (possibly infinite) t_0, t_1 such that the integral converges. Rewriting (2.20) with respect to conformal time, we obtain the alternative metric tensor, the *conformal FLRW metric*,

$$g = a(\tau)^2 \big(-\mathrm{d}\tau \otimes \mathrm{d}\tau + \mathrm{d}r \otimes \mathrm{d}r + f_K(r)^2 (\mathrm{d}\theta \otimes \mathrm{d}\theta + \sin^2\theta\,\mathrm{d}\phi \otimes \mathrm{d}\phi)\big) \qquad (2.21)$$

and thus we notice that a flat FLRW universe is locally conformally isometric to Minkowski space.

Throughout this thesis we mostly work with flat FLRW universes and will fix $M \simeq \mathbb{R}^4$ in that case. Therefore we can choose globally Cartesian coordinates for the spatial sections so that we have

$$g = -\mathrm{d}t \otimes \mathrm{d}t + a(t)^2 \delta_{ij}\,\mathrm{d}x^i \otimes \mathrm{d}x^j = a(\tau)^2 \big(-\mathrm{d}\tau \otimes \mathrm{d}\tau + \delta_{ij}\,\mathrm{d}x^i \otimes \mathrm{d}x^j\big),$$

where the coordinate functions x^i range over the entire real line.

2.3.2 Gauge problem

The correspondence of a *background spacetime* (M, g_0) to the *physical spacetime* (M, g) is equivalent to the specification of a diffeomorphism $\psi : M \to M$. Given any other diffeomorphism ψ' between the background spacetime and the physical spacetime, we can construct $\varphi = \psi^{-1} \circ \psi'$. For a tensor field S on M to be *gauge invariant* we require that $\varphi^* S = S$. Then it holds that

$$\delta S = S - \psi^* S = S - \psi'^* S$$

and we say that the perturbation δS is a gauge invariant quantity. Otherwise, the perturbation δS is completely dependent on the mapping ψ and even if ψ is specified it will not be an observable quantity unless the correspondence ψ itself has been specified via an observational procedure (*e.g.* via an averaging approach, see the

discussion in [79]). Therefore the only possibilities for a tensor field δS to be gauge invariant is that S is a constant scalar, a zero tensor or a product of Kronecker deltas [204, Lem. 2.2]).

A slightly different picture which clarifies the perturbation aspect of the gauge problem is sometimes more helpful. Since the local flow of any vector field v on M is a diffeomorphism $\psi_\varepsilon : M \to M$, where ε is contained in a sufficiently small interval around 0, we can describe the gauge problem alternatively in terms of vector fields. Given a (differentiable) tensor field S, we have by definition

$$\mathscr{L}_v S = \lim_{\varepsilon \to 0} \frac{1}{\varepsilon}(\psi_\varepsilon^* S - S)$$

or, in other words,

$$\psi_\varepsilon^* S = S + \varepsilon \mathscr{L}_v S + O(\varepsilon^2).$$

The gauge choice is now encoded in v, which is completely arbitrary, and we see that S is gauge invariant to first order if only if $\mathscr{L}_v S = 0$ for all v. For S to be exactly gauge invariant (as discussed above) it must hold that $\mathscr{L}_v^n S = 0$ for all n.

In the light of this discussion, two approaches to perturbations of FLRW spacetimes seem expedient: Since any quantity describing the inhomogeneity or anisotropy of the perturbation of the FLRW spacetime must vanish on the background, it will be gauge-invariant. This leads to the "1 + 3 covariant and gauge-invariant" approach of [79, 119, 147]. The alternative and more commonly used approach due to [25, 140] constructs gauge invariant quantities directly from the perturbed metric and stress-energy tensor.

2.3.3   Decomposition of tensor fields

Before we can discuss metric perturbations, we need to investigate the decomposition of vector and rank-2 tensor fields into their 'scalar', 'vector' and 'tensor' parts [203].

Consider a non-compact,[4] boundaryless, orientable 3-dimensional Riemannian manifold (Σ, γ) with covariant derivative denoted by a vertical bar (*e.g.*, $\phi_{|i}$). Using the Hodge decomposition theorem, we can uniquely decompose any sufficiently fast decaying smooth one-form B as

$$B_i = \phi_{|i} + S_i, \tag{2.22}$$

where ϕ is a scalar function and S_i is divergence-free.

Any $(0, 2)$ tensor field C can be decomposed as

$$C_{ij} = \tfrac{1}{3}\delta_{ij}\gamma^{kl}C_{kl} + C_{[ij]} + C_{\langle ij \rangle},$$

i.e., into its trace, its antisymmetric part (equivalent to a vector field via the Hodge operator $*$) and its trace-free symmetric part $C_{\langle ij \rangle}$. According to [53, Thm. 4.3], as a consequence of the Fredholm alternative and an application of (2.22), a sufficiently fast decaying C can be further uniquely decomposed so that

$$C_{\langle ij \rangle} = \phi_{|\langle ij \rangle} + S_{(i|j)} + h_{ij}, \tag{2.23}$$

where ϕ and S are as before and h is a trace- and divergence-free $(0, 2)$ tensor field.

---

[4]The decomposition works also for compact manifold but is non-unique in that case.

2.3.4 Metric perturbations

As background spacetime we take (M, g_0) with g_0 the flat FLRW metric in conformal time, *i.e.*,

$$g_0 = a^2(-\mathrm{d}\tau \otimes \mathrm{d}\tau + \delta_{ij}\,\mathrm{d}x^i \otimes \mathrm{d}x^j),$$

which will be used throughout this section to raise and lower indices. Then we define the *perturbed FLRW metric*

$$
\begin{aligned}
a^{-2}g = {} & -(1 + 2\phi)\,\mathrm{d}\tau \otimes \mathrm{d}\tau + B_i\,(\mathrm{d}x^i \otimes \mathrm{d}\tau + \mathrm{d}\tau \otimes \mathrm{d}x^i) \\
& + \big((1 - 2\psi)\delta_{ij} + 2C_{ij}\big)\,\mathrm{d}x^i \otimes \mathrm{d}x^j
\end{aligned}
\tag{2.24}
$$

where the scalar fields ϕ, ψ, the 3-vector field B and the trace-free 3-tensor field C are considered 'small', *i.e.*, (2.24) should be understand as a 1-parameter family of metrics g_ε and each of the perturbation variables as multiplied with a small parameter ε. However, to avoid cluttering the equations unnecessarily, one usually omits the ε.

The ten degrees of freedom encoded in these four quantities exhibit the full gauge dependence. Before studying the behaviour of g under gauge transformations, let us rewrite (2.24) using the decompositions (2.22) and (2.23):

$$B_i = B_{,i} - S_i, \quad C_{ij} = E_{,ij} + F_{(i,j)} + \tfrac{1}{2}h_{ij}$$

with two scalar fields E, B, two divergence-free 3-vector fields S, F and a trace-free, transverse 3-tensor field h. Therefore,

$$
\begin{aligned}
a^{-2}g = {} & -(1 + 2\phi)\,\mathrm{d}\tau \otimes \mathrm{d}\tau + (B_{,i} - S_i)(\mathrm{d}x^i \otimes \mathrm{d}\tau + \mathrm{d}\tau \otimes \mathrm{d}x^i) + \\
& + \big((1 - 2\psi)\delta_{ij} + 2E_{,ij} + 2F_{(i,j)} + h_{ij}\big)\,\mathrm{d}x^i \otimes \mathrm{d}x^j.
\end{aligned}
\tag{2.25}
$$

This decomposition allows us to consider three types of perturbations separately, namely, the *scalar perturbations* caused by ϕ, ψ, E, B, the *vector perturbations* due to S, F and the *tensor perturbations* caused by h. The inverse of the perturbed metric (2.24) up to first order is

$$a^2 g^{-1} = -(1 - 2\phi)\,\partial_\tau \otimes \partial_\tau + B^i\,(\partial_i \otimes \partial_\tau + \partial_\tau \otimes \partial_i) + \big((1 + 2\psi)\delta^{ij} - 2C^{ij}\big)\,\partial_i \otimes \partial_j.$$

Let us now determine the transformation behaviour of these perturbation variables by calculating the gauge dependence of g up to linear order, *i.e.*, the dependence of $\mathscr{L}_\xi g_{ab} = 2\nabla_{(a}\xi_{b)} = 2\,{}^0\nabla_{(a}\xi_{b)} + O(\varepsilon)$ on $\xi^\mu = (\xi^0, \xi^{,i} + \xi^i)$ with $\xi^i{}_{,i} = 0$, where ${}^0\nabla$ is the covariant derivative on the background spacetime. Hence we have the transformations:

$$
\begin{aligned}
g_{00} &\to g_{00} - 2a^2(\xi^{0\prime} + \mathscr{H}\xi^0), \\
g_{0i} &\to g_{0i} + a^2(\xi'_{,i} + \xi'_i - \xi^0{}_{,i}), \\
g_{ij} &\to g_{ij} + 2a^2(\xi_{,(ij)} + \xi_{(i,j)} + \mathscr{H}\delta_{ij}\xi^0),
\end{aligned}
$$

where we define the *conformal Hubble parameter* $\mathscr{H} \doteq aH = a'/a$ and a prime denotes a derivative with respect to the conformal time (*e.g.*, $a' = \partial_\tau a$). It follows that the

perturbation variables transform as

$$
\begin{aligned}
\phi &\to \phi + \xi^{0\prime} + \mathcal{H}\xi^0 & \psi &\to \psi - \mathcal{H}\xi^0 \\
B &\to B + \xi' - \xi^0 & E &\to E + \xi \\
S_i &\to S_i - \xi'_i & F_i &\to F_i + \xi_i
\end{aligned}
$$

and $h_{ij} \to h_{ij}$.

Constructing linear combinations of the perturbation variables, we can now construct several first-order gauge invariant quantities. Two simple (first-order) gauge-invariant functions characterizing the scalar perturbations are the *Bardeen potentials* Φ, Ψ [25]

$$
\Phi \doteq \phi - \mathcal{H}\sigma - \sigma', \tag{2.26a}
$$

$$
\Psi \doteq \psi + \mathcal{H}\sigma \tag{2.26b}
$$

in terms of the shear potential $\sigma \doteq E' - B$.

3

Analysis

Summary

The contents of the present chapter may be generously subsumed under "analysis", hence the title. We begin with a review of topological vector spaces (Sect. 3.1) starting from the basics of topology and locally convex topological spaces with the particular example of function spaces, and finishing with a discussion duality pairings and tensor products. The material presented in this section is a excerpt of the results found in the books [133, 137, 176, 195, 211]. As always, although we refer the reader to the books cited above for proofs of the various statements, care has been taken to present the material in a structured way so that no results should appear surprising.

The second section (Sect. 3.2) concerns the theory of $*$-algebras and thus forms the foundation of the algebraic approach to quantum field theory to be discussed in the later chapters. Here we will discuss the general features of $*$-algebras and C^*-algebras, states with an emphasize on the Gel'fand–Naimark–Segal reconstruction theorem, and the Weyl C^*-algebra. For this section we refer the reader to the books [99, 130]. Details on the Weyl algebra can be found in [24, 151, 158, 200].

Functional derivatives have always played an important role in physics but on infinite dimensional spaces, which appear naturally in quantum field theory, they are very subtle. We introduce two different notions of functional derivatives in the third section (Sect. 3.3): the directional, or Gâteaux, derivative and the Fréchet derivative. We will show that the two derivatives are closely related. Proofs and more information on the directional derivative can be found in [116, 163].

In the fourth section (Sect. 3.4) we will pick apart the Banach fixed-point theorem and prove several statements on the existence and uniqueness of fixed-points and their properties. These results will form the basis of the proof existence of solutions to the semiclassical Eisntein equation to be presented in Chap. 8. Some of the results presented in the fourth section are already shown in [3] by Pinamonti and the author.

The theory of distributions plays a fundamental role in quantum field theory: quantum fields can be understood as distributions. Therefore, we will discuss in detail distributions and microlocal analysis in the fifth section (Sect. 3.5). That is, we will review the basic definition of distributions and distributional sections on manifolds, the nuclearity property of the associated function spaces and the Schwartz kernel theorem, the Fourier transform and Schwartz functions and distributions, the wavefront set of distribution and distributional sections, including its behaviour under various operations such as pullbacks, and finally the important propagation of singularities theorem. Good references for this section are the books by Hörmander [125–128] and also [207, 211]. Several recent results on the properties of spaces of distributions may be found in [54]. An excellent introduction to the wavefront set with several examples is [41] by the same author.

In the last section of this chapter (Sect. 3.6) we discuss wave equations. Since bosonic quantum field typically satisfy a wave equation and the Dirac-type equations satisfied by fermionic quantum fields are closely related, an understanding of the solutions of these equations is very important. With some minor modifications we mostly follow [23, 24] to introduce the advanced, retarded and causal propagators of normally and pre-normally hyperbolic differential operators and their relation to the Cauchy problem. Several extensions of the results in [23, 24], that are also partially stated here, can be found in [160, 220].

3.1   Topological vector spaces

3.1.1   Topology

A *topological space* is a set X of points with a notion of neighbourhoods. More concretely, besides X it consists of a collection τ of subsets, the *open sets*, such that

  (a) both $\emptyset$ and X are open,

  (b) the union of any collection of open sets is open,

  (c) any finite intersection of open sets is open.

We call the collection τ the *topology* of X; examples are illustrated in Fig. 3.1.

There are two topologies that can be defined for every set. The *discrete topology* of a set contains all its subsets, whereas the *trivial topology* consists only of the empty set and the set itself.

Given two topologies τ and τ' on the same set, we can compare them: If $\tau \subset \tau'$, we say that τ' is *finer* than τ and that τ is *coarser* than τ'. It follows that for any set the trivial topology and the discrete topology are respectively the coarsest and finest possible topology.

The complements of the open sets are the *closed sets*. By the de Morgan laws, they have the following properties: both $\emptyset$ and X are close, the intersection of any collection of closed sets is closed and any finite union of closed sets is closed. It is possible for a set to be both closed and open or neither. If the only sets in X that are both open and closed are $\emptyset$ and X, then X is *connected*.

The *closure* $\mathrm{cl}\, U$ of a set $U \subset X$ is the intersection of all closed set that contain U. The subset U is called *dense* in X if its closure is X: $\mathrm{cl}\, U = X$.

A *neighbourhood* of a point $x \in X$ is an open set $U \in \tau$ that contains x. If for each pair of distinct points x, y in a topological space X there exist disjoint neighbourhoods U and V of x and y, then it is called a *Hausdorff space* (Fig. 3.1). Obviously, endowing a set with the discrete topology turns it into a Hausdorff space.

A *basis* (or *base*) of a topological space X is a collection $\mathfrak{B}$ of open sets in X such that every open set in X can be written as the union of elements of $\mathfrak{B}$; one can say that the topology of X is generated by $\mathfrak{B}$. If the basis of X is countable, we say that X is *second-countable*. A *local basis* $\mathfrak{B}(x)$ of a point $x \in X$ is defined as a collection of neighbourhoods of x such that every neighbourhood of x is a superset of an element of the local basis. The union of all local bases is a basis. If every point
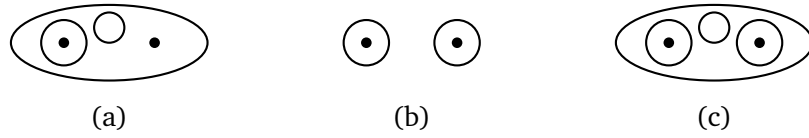
Figure 3.1. Collections of subsets of two points: (a) is a topology, (b) is not a topology (the empty set and the whole set are missing), (c) is the discrete topology (thus a Hausdorff topology).

has a countable basis, we say that X is *first-countable*. Clearly, a second-countable space is first-countable but the implication cannot be reversed.

We say that a map $f : X \to Y$ between two topological spaces X, Y is *continuous* if for all $x \in X$ and all neighbourhoods V of $f(x)$ there is a neighbourhood U of x such that $f(U) \subset V$. The *space of all continuous maps* between X and Y is denoted $C(X, Y)$ or $C^0(X, Y)$ and by $C(X) = C^0(X)$ if $Y = \mathbb{R}$. A bijective map between two topological spaces X, Y is a *homeomorphism* is both f and f^{-1} are continuous. X and Y are then called *homeomorphic*, *i.e.*, they are topologically equivalent.

Let X be a set and $Y_i, i \in I$ a family of topological spaces with topologies τ_i. Given maps $f_i : X \to Y_i$, the *initial topology* on X is the *coarsest* topology such that the f_i are continuous. It is generated by the finite intersections of $\{f_i^{-1}(U) \mid U \in \tau_i\}$. Examples of the initial topology are the *subspace topology*, *i.e.*, the topology induced on a subspace $X \subset Y$ by the inclusion map $\iota : X \hookrightarrow Y$, and the *product topology*, *i.e.*, the topology induced on a product space $X = \prod_{i \in I} Y_i$ by the projections $\pi_i : X \to Y_i$. Conversely, given maps $f_i : Y_i \to X$, the *final topology* on X is the *finest* topology such that each f_i is continuous. It is given as

$$\tau = \{U \subset X \mid f_i^{-1}(U) \in \tau_i \ \forall i \in I\}.$$

Important examples are the *quotient topology* on a quotient space $X = Y/\sim$ with the map given by the canonical projection $Y \to Y/\sim$ and the *direct sum topology* on the direct sum $X = \sum_{i \in I} Y_i$ given by the canonical injections $Y_i \to X$.

Another application for the initial topology is the topology induced by a pseudo-metric: A *pseudometric* on a set X is a map $\mapsto d(\cdot, \cdot) : X \times X \to \mathbb{R}$ such that for all $x, y, z \in X$:

(a) $d(x, x) = 0$,

(b) $d(x, y) = d(y, x)$ (symmetry),

(c) $d(x, z) \le d(x, y) + d(y, x)$ (triangle inequality);

the set X together with d is a *pseudometric space*. If the pseudometric satisfies $d(x, y) > 0$ (positivity) for all $x \ne y$, then it is a *metric* and X is a *metric space*. The pseudometric on X induces the initial topology on X which is generated by the open balls around each point:

$$\mathfrak{B}(y) = \{x \in X \mid d(x, y) < r\}.$$

A map $f : X \to Y$ between two pseudometric spaces with pseudometrics d_X, d_Y is called an *isometry* if $d_X(x, y) = d_Y(f(x), f(y))$ for all $x, y \in X$.

Given a topological space X, a sequence of points (x_n) in X *converges* to a point $x \in X$ if every neighbourhood U of x contains all but finitely many elements of the sequence. If X is a pseudometric space, it is called *complete* if every Cauchy sequence with respect to its pseudometric converges to a point in X. Every pseudometric space X that is not complete can be completed. Namely, the *completion* of X is the (essentially unique) complete pseudometric space $\overline{X}$ of which X is a dense isometric subspace.

A topological (sub)space X is called *compact* if each of its *open covers*, *viz.*, a collection of subsets whose union contains X as a subset, has a finite subcollection that also contains X. If X is compact, then every subset $Y \subset X$ is also compact in the subspace topology. The space X is *locally compact* if every point in X has a compact neighbourhood. Furthermore, X is called *σ-compact* if it is the union of countably many compact subsets. Any compact space is locally compact and σ-compact; the converse is, however, false. One can also show that every second-countable and locally compact Hausdorff space (thus, in particular, every topological manifold) is σ-compact. We say that a map $f : X \to Y$ between two topological spaces X, Y is *proper* if the preimage of every compact set in Y is compact in X.

A related concept is that of boundedness. A subset $U \subset X$ of a pseudometric space X is *bounded* if for each $x, y \in U$ there exists a r such that $d(x, y) \leq r$.

3.1.2 Locally convex topological vector spaces

A *topological vector space* is a $\mathbb{K}$-vector space X such that the vector space operations of addition and scalar multiplication are (jointly) continuous with respect to the topology of X. Since addition is continuous and the topology therefore translation-invariant, the topology of X is completely determined by the local basis $\mathfrak{B}(0)$ at the origin. For the same reason one can show that X is Hausdorff if and only if $\{0\}$ is closed. Note that in a Hausdorff topological vector space every complete subset is also closed. The (topological) vector space X is called *convex* if $x, y \in X$ implies $\lambda x + (1 - \lambda)y \in X$ for all $\lambda \in [0, 1]$.

Let X, Y be topological vector spaces and $W \subset X$ a subset. A mapping $f : W \to Y$ is *uniformly continuous* if to every neighbourhood V of zero in Y there exists a zero neighbourhood $U \subset X$ such that for all $x, y \in W$

$$x - y \in U \implies f(x) - f(y) \in V.$$

Every uniformly continuous map is already continuous but the converse is not true. However, if f is linear and also W is a vector subspace, then continuity also implies uniform continuity. Moreover, if W is a dense subset, then to every uniformly continuous map $f : W \to Y$ there exists a unique continuous map $\overline{f} : X \to Y$ which extends f. The extension of a linear map from a dense vector subspace is even uniformly continuous and linear.

It is usually desirable to have topological vector spaces with additional structures: A *seminorm* on a $\mathbb{K}$-vector space X is a map $\| \cdot \| : X \to \mathbb{R}$ such that for all $x, y \in X$ and $\lambda \in \mathbb{K}$:

(a) $\|x\| \geq 0$ (positive-semidefiniteness),

(b) $\|\lambda x\| = |\lambda|\|x\|$ (absolute homogeneity),

(c) $\|x + y\| \leq \|x\| + \|y\|$ (triangle inequality).

If the seminorm satisfies $\|x\| > 0$ (positivity) for all $x \neq 0$, then it is a *norm* and the vector space is called *normed space*. Note that each (semi)norm induces a (translation-invariant) (pseudo)metric $d(x, y) = \|x - y\|$.

Analogously to the pseudometrics, given a *family of seminorms* $\{\|\cdot\|_i\}_{i \in I}$ on a vector space X, they induce the initial topology on X. More explicitly, the topology is generated by all finite intersections of $\{x \in X \mid \|x\|_i < r\}$, the open balls around the origin.[1] If I is countable, we can assume that $I \subset \mathbb{N}$ and the topology above is the same topology as the one induced by the metric

$$d(x, y) = \sum_{k \in I} \frac{1}{2^k} \frac{\|x - y\|_k}{1 + \|x - y\|_k},$$

where the factors 2^{-k} may be replaced by the coefficients of any convergent series.

We say that a family of seminorms $\{\|\cdot\|_i\}$ on a vector space X is *separating* if for every nonzero $x \in X$ there exists an i such that $\|x\|_i > 0$. It is immediate that a vector space with topology induced by a family of seminorms is Hausdorff if and only if the seminorms are separating.

Let X be vector space endowed with a family of seminorms $\{\|\cdot\|_i\}_{i \in I}$. Then we can define the following topological vector spaces in order of generality:

1. if the family of seminorms is separating, then X is a *locally convex (topological vector) space*;

2. if, in addition, I is countable and X is complete with respect to each of its seminorms, then X is a *Fréchet space*;

3. if, in addition, the family of seminorms consists of only one *norm*, then X is a *Banach space*.

Finally note that on all these spaces above the *Hahn–Banach theorem* can be applied. That is, given a $\mathbb{K}$-vector space X with a seminorm $\|\cdot\|$ and a linear form $f : U \to \mathbb{K}$ on a vector subspace U such that $|f(x)| \leq \|x\|$ for all $x \in U$, there exists a (generally non-unique) linear form $\overline{f} : X \to \mathbb{K}$ which extends f such that $|\overline{f}(x)| \leq \|x\|$.

3.1.3 Topologies on function spaces

Important vector spaces are subspaces of the *space of functions*

$$Y^X \doteq \{f : X \to Y\}$$

between a set X and a topological space Y. Y^X can be equipped with the *topology of pointwise convergence*, which is just the product topology with the projections

---

[1]Conversely, given a basis of the origin it is possible to construct a family of seminorms from its elements.

$\pi_x : Y^X \to Y, f \mapsto f(x)$. In this topology a sequence of functions (f_n) converges to some f if and only if each $f_n(x)$ converges to $f(x)$ at all x.

If (Y, d) is a pseudometric space, the space of functions can be equipped another topology, the *topology of uniform convergence*. In this case, a subspace $Z \subset Y^X$ can be endowed with the pseudometric

$$d(f, g) = \sup_{x \in X} d\big(f(x), g(x)\big)$$

for all $f, g \in Z$, which induces a topology for Z. Note, however, that Z with this topology is not really a topological vector space because multiplication will fail to be continuous unless Z is a subset of the bounded functions. A sequence of functions (f_n) converges to some f if and only if for every $\varepsilon > 0$ there exists a N such that $d(f_n, f) < \varepsilon$ for all $n \geq N$.

If X is a topological space, yet another topology on subspaces $Z \subset Y^X$ is the *compact-open topology*. For all compact $K \subset X$ and open $U \subset Y$ it is generated by the finite intersections of

$$\{f \in Z \mid f(K) \subset U\},$$

i.e., the set of functions that carry compact subsets into open subsets. The compact-open topology is finer than the topology of pointwise convergence. If (Y, d) is a pseudometric space, the compact-open topology is the initial topology induced by the pseudometric on compact subsets, *i.e.*, it is generated by the finite intersections of

$$\{f \in Z \mid \sup_{x \in K} d\big(f(x), 0\big) < r\}.$$

Therefore it is also called the *topology of uniform convergence in compacta* and, if X is compact, it is the same as the topology of uniform convergence. It follows that a sequence of functions (f_n) converges to some f if and only if for every $\varepsilon > 0$ and compact $K \subset X$ there exists a N such that $d(f_n(x), f(x)) < \varepsilon$ for all $n \geq N$ and $x \in K$.

Again, suppose that X, Y are topological vector spaces and $Z \subset Y^X$. Z is called *equicontinuous* if for every neighbourhood of the origin $U \subset X$ there exists a neighbourhood of the origin $V \subset X$ such that $f(U) \subset V$ for every $f \in Z$. Equicontinuity for a set of one element is of course the same as continuity.

3.1.4 Duality

A *duality* or *dual pairing* $\langle Y, X \rangle$ is a triple $(X, Y, \langle \cdot, \cdot \rangle)$ of two vector spaces X, Y and a *non-degenerate* bilinear form $\langle \cdot, \cdot \rangle : Y \times X \to \mathbb{K}$, *i.e.*,

$$\langle y, x \rangle = 0 \text{ for all } y \in Y \text{ implies } x = 0$$
$$\langle y, x \rangle = 0 \text{ for all } x \in X \text{ implies } y = 0.$$

The standard example of a duality is that between a vector space X and its algebraic dual X^*, where the pairing is given by the *canonical bilinear form*

$$\langle \cdot, \cdot \rangle : X^* \times X \to \mathbb{K}, (f, x) \mapsto \langle f, x \rangle \doteq f(x).$$

A more important example is that of the *topological dual* $X' \subset X^*$ of a topological space X, which consists of all continuous linear maps. Note that the pairing $\langle X', X \rangle$ is not a proper duality unless X is Hausdorff because the restriction of the canonical bilinear form to $X' \times X$ is non-degenerate if and only if X is Hausdorff.

For each $x \in X$ the map $x \mapsto \langle y, x \rangle$ gives an injective map of X into Y^* and an analogous construction embeds Y into X^*. In the following, the identification of X with a subspace of Y^* and Y with a subspace of X^* will always be tacitly assumed unless otherwise noted.

In particular, X is a subspace of $\mathbb{K}^Y$ and can therefore be equipped with pointwise topology. This locally convex Hausdorff topology is called the *weak topology* on X with respect to $\langle Y, X \rangle$; statements for weak topology will often be indicated by the adjective "weakly". The weak topology is the coarsest topology such that $x \mapsto \langle y, x \rangle$ is continuous for all $y \in Y$ and one finds $Y = X'$ with respect to the weak topology on X. Moreover, if Y is locally convex, the seminorms on Y yield *dual seminorms* on X given by

$$\|x\|_i \doteq \sup \left\{ |\langle y, x \rangle| \,\middle|\, y \in Y \text{ with } \|y\|_i \leq 1 \right\},$$

which also induce the weak topology.

The statements above can also be made with the role of X and Y interchanged to introduce the weak topology on Y with respect to $\langle Y, X \rangle$. In particular, if $Y = X'$ and $\langle \cdot, \cdot \rangle$ the canonical bilinear form, then X' with the weak topology is called the *weak dual*. Furthermore, given a subset $Y \subset X^*$, then the induced pairing between Y and X is non-degenerate if and only if Y is weakly dense in X^*. Thus any Y is weakly complete if and only if $Y = X^*$.

Another (locally convex Hausdorff) topology on X induced by a duality $\langle Y, X \rangle$ with a locally convex space Y is the *strong topology*, which is the topology of uniform convergence on the bounded subsets of Y; statements for strong topology will be often be indicated by the adjective "strongly". It is induced by the family of seminorms

$$\|x\|_B = \sup_{y \in B} |\langle y, x \rangle|$$

for each bounded set $B \subset Y$. Again, we can interchange the role of X and Y and call X' endowed with the strong topology induced by the canonical pairing the *strong dual*.

If we equip X with the strong topology with respect to $\langle Y, X \rangle$, then the map $x \mapsto \langle y, x \rangle$ will not be continuous for any $y \in Y$. The finest topology on X such that this map is continuous is called the *Mackey topology* but it will not concern us here any further.

Finally, note that the dual of a Banach space is *always* a Banach space, but the dual of a Fréchet space that is not Banach is *never* a Fréchet space.

3.1.5   Tensor products on locally convex spaces

Given two locally convex topological vector spaces X, Y there are many different ways to define a family of seminorms for the space $X \otimes Y$. Therefore there is no natural topology for $X \otimes Y$ if X or Y is infinite-dimensional, whence one speaks of different

topological tensor products. The most common topological tensor products are the projective and injective tensor product introduced below.

The *projective tensor product topology* equips the algebraic tensor product $X \otimes Y$ with the finest topology such that $\otimes : X \times Y \to X \otimes Y$ is *jointly* continuous. That is, it is the final topology defined by the projections $\pi_X : X \to X \otimes Y$, $\pi_Y : Y \to X \otimes Y$. Equivalently, the topology is induced by the seminorms

$$\|u\|_{i,j} = \inf \left\{ \sum\nolimits_k \|x_k\|_i \|y_k\|_j \,\Big|\, u = \sum\nolimits_k x_k \otimes y_k \right\}$$

for all $u \in X \otimes Y$ and where the infimum runs over all representations of u. The resulting locally convex space is usually denoted $X \otimes_\pi Y$ and its completion $X \widehat{\otimes}_\pi Y$.

A coarser topology is defined by the *injective tensor product topology*; it is the finest topology such that $\otimes : X \times Y \to X \otimes Y$ is *separately* continuous. Let X', Y' be the weak duals of X, Y. Note that $X \otimes Y$ can be embedded into the space of bilinear *separately* continuous maps $X' \times Y' \to \mathbb{K}$, denoted $B(X, Y)$, with the topology of uniform convergence $U \times V$ on all equicontinuous sets $U \subset X$ and $V \subset Y$. That is, the topology is generated for all $U \times V$ and open $I \subset \mathbb{K}$ by the finite intersections of

$$\{f \in B(X, Y) \mid f(U \times V) \subset I\}.$$

$X \otimes Y$ can now be endowed with the corresponding subspace topology. Seminorms that induce this topology are given by

$$\|u\|_{i,j} = \sup \left\{ |(f \otimes g)(u)| \,\Big|\, f \in X', g \in Y' \text{ such that } \|f\|_i = \|g\|_j = 1 \right\}.$$

The space $X \otimes Y$ equipped with the injective topology is usually denoted $X \otimes_\varepsilon Y$ and its completion is denoted $X \widehat{\otimes}_\varepsilon Y$. Observe that $C(X, Y) \simeq C(X) \widehat{\otimes}_\varepsilon Y$ if Y is complete.

Locally convex spaces on which the injective and projective tensor product agree are called nuclear. More precisely, we say that a locally convex space X is *nuclear* if

$$X \otimes_\varepsilon Y = X \otimes_\pi Y \quad \text{or, equivalently,} \quad X \widehat{\otimes}_\varepsilon Y = X \widehat{\otimes}_\pi Y$$

for every locally convex space[2] Y in which case we simply write $X \otimes Y$. If both X and Y are nuclear, then also $X \otimes Y$ is nuclear. Moreover, if a subspace of a nuclear space is nuclear and the quotient space of a nuclear space by a closed subspace is nuclear.

A more useful characterisation of nuclear spaces is in terms of summable sequences. Denote by $\ell_1(X)$ the X-valued *summable sequences*, i.e., the set of sequences (x_n) in X such that all unordered partial sums $\sum_{n \in I \subset \mathbb{N}} x_n$ converge in X. Further, denote by $\ell_1\{X\}$ the X-valued *absolutely summable sequences*, i.e., the set of sequences (x_n) in X such that $\sum_n \|x_n\|_i < \infty$ for all seminorms $\|\cdot\|_i$ of X. Then X is nuclear if and only if

$$\ell_1(X) = \ell_1\{X\}$$

and hence, by the above observation, both sides are equal to $\ell_1 \otimes X = \ell_1 \otimes_\varepsilon X = \ell_1 \otimes_\pi X$. In other words, X is nuclear if and only if every summable sequence in X is already absolutely summable. Nuclear spaces are therefore very similar to finite-dimensional

---

[2]Actually it is sufficient to check equality for $Y = \ell_1$, see below.

spaces and, while *every* finite-dimensional locally convex space is nuclear, *no* infinite-dimensional normed space is.

We finish this section by stating what can be called the *abstract kernel theorem* for Fréchet spaces:

$$(X \otimes Y)' \simeq X' \otimes Y' \quad \text{and} \quad (X \widehat{\otimes} Y)' \simeq X' \widehat{\otimes} Y'$$

for every X, Y such that X (or Y) is nuclear and where all duals are strong duals.

3.2   Topological *-algebras

A *topological *-algebra* is a topological algebra $\mathcal{A}$ over $\mathbb{C}$, *i.e.*, a topological $\mathbb{C}$-vector space with a separately continuous ring multiplication, together with a continuous *involution* *. That is, there is an automorphism

$$* : \mathcal{A} \to \mathcal{A}, \quad x \mapsto x^*,$$

which is antilinear and involutive such that

(a) $(ax + by)^* = \bar{a}x^* + \bar{b}y^*$,

(b) $(xy)^* = y^*x^*$,

(c) $(x^*)^* = x$

for all $x, y \in \mathcal{A}$ and $a, b \in \mathbb{C}$. If, in addition, $\mathcal{A}$ has a multiplicative unit $\mathbf{1}$, we say that $\mathcal{A}$ is a *unital *-algebra*. Elements x, y of the algebra $\mathcal{A}$ are called

$$
\begin{aligned}
\textit{adjoint} \quad &\text{if} \quad x^* = y, \\
\textit{self-adjoint} \quad &\text{if} \quad x^* = x, \\
\textit{normal} \quad &\text{if} \quad x^*x = xx^*, \\
\textit{unitary} \quad &\text{if} \quad x^*x = \mathbf{1} = xx^*,
\end{aligned}
$$

where unitarity obviously requires the existence of a unit element. Note that $\mathbf{1}$ is always self-adjoint.

A *-subalgebra $\mathcal{I} \subset \mathcal{A}$ is called a left (right) *-ideal* if yx (resp. xy) is in $\mathcal{I}$ for all $y \in \mathcal{I}$ and $x \in \mathcal{A}$. If the subalgebra is both a left and right *-ideal, it is just called a (two-sided) *-ideal. It follows that an ideal $\mathcal{I}$ of $\mathcal{A}$ is a *-ideal if and only if $\mathcal{I}^* = \mathcal{I}$.

The homomorphisms that arise between *-algebras, called *-*homomorphisms*, are those that preserve in addition to the multiplicative also the involutive structure, *i.e.*, a map $\alpha : \mathcal{A} \to \mathcal{B}$ is a *-homomorphisms if it is an algebra-homomorphism and $\alpha(x^*) = \alpha(x)^*$ for all $x \in \mathcal{A}$. If the *-algebras are unital, we also demand that *-homomorphisms be unit-preserving.

Often one needs a *-algebra which also has the structure of a normed vector space. In the case of *-algebras, it makes sense to require the norm to satisfy an additional property: A norm $\| \cdot \| : \mathcal{A} \to \mathbb{R}$ is said to be a C^*-*norm* if

$$\|x^*x\| = \|x\|^2$$

for all $x \in \mathcal{A}$. This differs from some definitions of C^*-norms because, in fact, every C^*-norm is automatically a *-isomorphism and submultiplicative, *i.e.*,

$$\|x^*\| = \|x\| \quad \text{and} \quad \|xy\| \leq \|x\|\|y\|$$

for all $x, y \in \mathcal{A}$ [199].

   If a *-algebra $\mathcal{A}$ comes equipped with such a C^*-norm $\|\cdot\|$ that turns $\mathcal{A}$ into a Banach space, then it is called a C^*-*algebra*. In a C^*-algebra, ring multiplication and inversion are continuous operations with respect to the norm; the continuity of addition, scalar multiplication and involution are obvious. The condition for a unital C^*-algebra to have a C^*-norm imposes such strong conditions on its algebraic structure that the algebra uniquely determines the norm. Namely,

$$\|x\|^2 = \|x^*x\| = \sup \left\{ |\lambda| \,\big|\, x^*x - \lambda\mathbf{1} \text{ is not invertible} \right\}$$

for every $x \in \mathcal{A}$. A *-homomorphism $\alpha : \mathcal{A} \to \mathcal{B}$ between two unital C^*-algebras is thus always norm-decreasing: $\|\alpha(x)\| \leq \|x\|$.

3.2.1   States

Given a *-algebra $\mathcal{A}$, one can consider its algebraic dual, the space of linear functionals on $\mathcal{A}$. A linear functional $\omega \in \mathcal{A}^*$ on a unital *-algebra $\mathcal{A}$ is *positive* if it satisfies

$$\omega(x^*x) \geq 0$$

for all $x \in \mathcal{A}$. If, $\mathcal{A}$ is unital and $\omega(\mathbf{1}) = 1$, we say that ω is *normalized*. A functional ω that is both positive and normalized is called a *state*. If the *-algebra $\mathcal{A}$ comes equipped with a topology, we always assume that ω is continuous, *i.e.*, we consider the topological dual $\mathcal{A}'$ instead of the algebraic one; an algebraic state on a C^*-algebra is automatically continuous with respect to the C^*-norm.

   Given positive $\omega \in \mathcal{A}'$, it follows that for all $x, y \in \mathcal{A}$

$$\omega(x^*y) = \overline{\omega(y^*x)},$$
$$|\omega(x^*y)|^2 \leq \omega(x^*x)\omega(y^*y),$$

where the second line is called the *Cauchy–Schwarz inequality*. If $\mathcal{A}$ is unital, the first equation implies that every positive ω is hermitian: $\omega(x^*) = \overline{\omega(x)}$.

   A state ω is *pure* if every other state η that is majorized by it, $\omega(x^*x) \geq \eta(x^*x)$, is of the form $\eta = \lambda\omega$ with $\lambda \in [0, 1]$. Consequently, a pure state cannot be written as the convex sum of two other states. States that are not pure are called *mixed*.

   The positive linear functionals equips $\mathcal{A}$ with a degenerate inner product via the antilinear pairing $\langle x, y \rangle = \omega(x^*y)$ which can be turned into a pre-Hilbert space by taking the quotient by the degenerate elements. This is the essential content of the famous *Gel'fand–Naimark–Segal construction*, usually abbreviated as *GNS construction*, which we will state after the following definition.

   A *-*representation* π of a *-algebra $\mathcal{A}$ is a *-homomorphism into the C^*-algebra of linear operators on a common dense (with respect to the norm $\|\cdot\| = \langle \cdot, \cdot \rangle^{1/2}$ on the Hilbert space) domain $\mathcal{D}$ of a Hilbert space $\mathcal{H}$. Note that the *-representation π is

continuous with respect to the *uniform operator topology*, *i.e.*, the topology induced by the *operator norm*

$$\|T\|_{\mathrm{op}} = \sup \left\{ \|Tx\| \,\middle|\, x \in \mathscr{D} \text{ with } \|x\| \leq 1 \right\}.$$

Moreover, if the domain $\mathscr{D}$ is complete in the *graph topology* induced by the family of seminorms $\|\cdot\|_x = \|\pi(x)\cdot\|$, we say the the *-representation π is *closed*.

   If there exists a vector $\Omega \in \mathscr{H}$ such that $\pi(\mathscr{A})\Omega = \{\pi(x)\Omega \mid x \in \mathscr{A}\}$ is dense in $\mathscr{H}$, then the *-representation is called *cyclic* and Ω *cylic vector*. If $\pi(\mathscr{A})\Omega$ is even dense in $\mathscr{D}$ in the graph topology, then π and Ω are called *strongly cyclic*. The state $\omega(x) = \langle \pi(x)\Omega, \Omega \rangle$ defined by a cyclic vector Ω of a cyclic *-representation π is *pure* if and only if the only subspaces left invariant by $\pi(A)$ are the trivial ones.

Theorem 3.1 (GNS construction). *Let ω be a state[3] on a unital topological *-algebra $\mathscr{A}$. Then there exists a closed (weakly continuous) strongly cyclic *-representation π of $\mathscr{A}$ on a Hilbert space $\mathscr{H}$ with inner product $\langle \cdot, \cdot \rangle$ and strongly cyclic vector Ω such that*

$$\omega(x^*y) = \langle \pi(x)\Omega, \pi(y)\Omega \rangle$$

for all $x, y \in \mathscr{A}$. The representation π is unique up to unitary equivalence.

   This is a standard theorem and a proof can be found in many places, *e.g.*, in [180]. Working with general *-algebras, we have not excluded the case of *-representations onto unbounded operators. For that reason it is not possible to uniquely extend the representation to the whole Hilbert space, and hence self-adjoint elements of the algebra might not be represented by self-adjoint operators but only symmetric operators. These problems do not occur if one applies the GNS construction to C^*-algebras.

3.2.2   Weyl algebra

Set V to be a $\mathbb{R}$-vector space and $\sigma : V \times V \to \mathbb{R}$ an antisymmetric bilinear form (*i.e.*, a pre-symplectic form).[4] A *Weyl *-algebra* $\mathscr{W}$ for (V, σ) is a unital involutive algebra generated by (nonzero) *Weyl generators* W, *i.e.*, symbols $W(\cdot)$ labelled by the vectors in V, which satisfy, for all $v, w \in V$, the relations

   (a) $W(v)W(w) = \exp\left(\frac{\mathrm{i}}{2}\sigma(v,w)\right)W(v+w)$,

   (b) $W(v)^* = W(-v)$.

Therefore the Weyl generators also have the following properties:

   (c) $W(0) = \mathbf{1}$,

   (d) $W(v)^* = W(-v) = W(v)^{-1}$,

---

[3]Actually it is enough for ω to be a positive; the normalization is not necessary for the theorem to hold.

[4]We follow [151] and will not assume that σ is non-degenerate. In fact it is sufficient to assume that σ is linear in its first *or* second argument.

  (e) $\big(W(v)\big)_{v \in V}$ are linearly independent.

Since $\mathscr{W}$ is generated by unitaries, every *-representation is necessarily by bounded operators. Moreover, between two Weyl *-algebras generated by Weyl generators W and W' for (V, σ) there exists a unique *-isomorphism α completely determined by $\alpha \circ W = W'$.

One can endow $\mathscr{W}$ with a C^*-norm, the *minimal regular norm*

$$\|x\| = \sup \big\{\omega(x^*x)^{1/2} \,\big|\, \omega \text{ is a state on } \mathscr{W}\big\}$$

for all $x \in \mathscr{W}$. If the bilinear form σ is non-degenerate, one can show that all C^*-norm over $\mathscr{W}$ are equal. We call the completion $\overline{\mathscr{W}}$ of a Weyl *-algebra $\mathscr{W}$, with respect to the minimal regular norm, the *Weyl C^*-algebra*. It is unique up to *-isomorphism and, in particular, *simple, viz.*, it has no non-trivial closed *-ideals, if and only if σ is non-degenerate [151].

The map $\mathbb{R} \ni \lambda \mapsto W(\lambda v)$ is not continuous in $\overline{\mathscr{W}}$, because $\|W(v) - W(w)\| = 2$ for all distinct $v, w \in V$ as a consequence of the spectral radius formula. A *-representation π of $\overline{\mathscr{W}}$ on a Hilbert space $\mathscr{H}$ is called *regular* if the one-parameter unitary groups

$$\lambda \mapsto (\pi \circ W)(\lambda v), \quad v \in V,$$

are strongly continuous. If the *-representation induced by a state on $\overline{\mathscr{W}}$ is regular, we also call the state regular. Invoking *Stone's theorem* $(\pi \circ W)(\lambda v)$, we can find a family of self-adjoint operators $F(\cdot)$ on $\mathscr{H}$, labelled by vectors in V, such that

$$(\pi \circ W)(\lambda v) = \exp\big(\mathrm{i}\lambda F(v)\big);$$

the map F is called the *field operator* and is generally unbounded.

A *strongly regular state* [23] is a regular state for which the operators $F(v)$, $v \in V$, have a common dense domain $\mathscr{D} \subset \mathscr{H}$, which is closed under the action of F, and for which $v \mapsto F(v)w$ is continuous for fixed $w \in V$. For strongly regular states the field operator is linear in its argument and thus a self-adjoint operator-valued distribution.

3.3   Derivatives

Various different notions of derivatives on topological vector spaces exist in the literature, see [20] for a survey and history of the topic. On infinite-dimensional spaces these derivatives are inequivalent and care must be taken to make precise which derivative is meant. On Banach spaces there exists the notable example of the Fréchet derivative. However, many spaces of interest in physics are not normed and so one must work with derivatives on more general spaces. Below we will define a directional derivative in the sense of Gâteaux and later compare it with the Fréchet derivative on Banach spaces.

3.3.1   The directional derivative

Let X, Y be two topological vector spaces and $U \subset X$ open. The *(directional) derivative* of a function $f : U \to Y$ at $x \in U$ in the direction $h \in X$ is defined as the map

$$\mathrm{d}f : U \times X \to Y,$$

$$\mathrm{d}_h f(x) \doteq \mathrm{d}f(x;h) \doteq \lim_{\varepsilon \to 0} \frac{1}{\varepsilon}\left(f(x + \varepsilon h) - f(x)\right) = \frac{\mathrm{d}}{\mathrm{d}\varepsilon} f(x + \varepsilon h)\bigg|_{\varepsilon=0}, \qquad (3.1)$$

if the limit exists. In particular, if f is a continuous linear function, then its derivative is $\mathrm{d}f(x;h) = f(h)$.

Note that the nomenclature here follows that of [116, 163] and differs from that in [3], where Pinamonti and the author called the same derivative *Gâteaux derivative*. The reason for this choice of a more neutral name is that the name "Gâteaux derivative" has sometimes been used for slightly different derivatives. However, all definitions known to the author agree whenever the derivative is both linear and continuous in the direction of the derivative.

It should be clear from the definition of the directional derivative, that the ordinary and partial derivative are special cases of the directional derivative for functions from Euclidean $\mathbb{R}^n$ to $\mathbb{R}$ or $\mathbb{C}$. Consequently the directional derivative is also closely related to the local form of the covariant derivative given by a connection on a vector bundle.

The function f is called *differentiable at x* if the limit exists for all $h \in X$ and simply *differentiable* if it is differentiable at every $x \in U$. Moreover, f is *continuously differentiable* or C^1 on U if the map $\mathrm{d}f$ is continuous (in the induced topology on $U \times X$). Higher derivatives may be defined recursively by

$$\mathrm{d}_{h_n} \cdots \mathrm{d}_{h_1} f(x) \doteq \mathrm{d}^n_{h_1,\dots,h_n} f(x) \doteq \lim_{\varepsilon \to 0} \frac{1}{\varepsilon}\left(\mathrm{d}^{n-1}_{h_1,\dots,h_{n-1}} f(x + \varepsilon h_n) - \mathrm{d}^{n-1}_{h_1,\dots,h_{n-1}} f(x)\right)$$

and we say that f is C^n if $\mathrm{d}f$ is C^{n-1}; if f is C^n for all $n \in \mathbb{N}$, then we say that f is C^∞ or *smooth*.

Continuity of the derivative already implies many other properties if the involved vector spaces are locally convex. Hence, let X, Y be locally convex spaces and $f : X \supset U \to Y$ be continuously differentiable. Then it can be shown that [116, 163]:

(a) the *fundamental theorem of calculus*

$$f(x + h) - f(x) = \int_0^1 \mathrm{d}f(x + th; h)\,\mathrm{d}t$$

holds if $x + [0,1]h \subset U$,

(b) f is locally constant if and only if $\mathrm{d}f = 0$,

(c) the map $h \mapsto \mathrm{d}f(x;h)$ is linear,

(d) f is continuous (not necessarily true if X, Y are not locally convex!),

(e) if $f \in C^n$, the map $(h_1,\dots,h_n) \mapsto \mathrm{d}^n f(x;h_1,\dots,h_n)$ is symmetric and multilinear and we can use yet another notation

$$\langle \mathrm{d}^n f(x), h_1 \otimes \cdots \otimes h_n \rangle \doteq \mathrm{d}^n f(x;h_1,\dots,h_n),$$

(f) if $f \in C^{n+1}$, *Taylor's formula*

$$f(x+h) = f(x) + \mathrm{d}f(x;h) + \cdots + \frac{1}{n!}\mathrm{d}^n f(x;h,\ldots,h)$$

$$+ \frac{1}{n!}\int_0^1 (1-t)^n \mathrm{d}^{n+1} f(x+th;h,\ldots,h)$$

holds for $x + [0,1]h \subset U$.

Moreover, given locally convex X, Y, Z, open subsets $U \subset X$, $V \subset Y$ and C^n maps $f : U \to V$, $g : V \to Z$, the *chain rule* holds for the composition $g \circ f$, i.e., also the composition $g \circ f$ is C^n [116, 163].

In ordinary calculus one can show that a continuously differentiable function function is locally Lipschitz. An analogous result holds for the directional derivative on normed spaces (see also [3]):

Proposition 3.2. *Let $f : X \to Y$ be a continuously differentiable map between the two normed spaces $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$. Then f is locally Lipschitz, i.e., for every convex neighbourhood U of $x_0 \in X$ there exists a $K \geq 0$ such that for all $x_1, x_2 \in U$*

$$\|f(x_1) - f(x_2)\|_W \leq K\|x_1 - x_2\|_V.$$

Proof. Since the derivative $\mathrm{d}f(x;h)$ is continuous and linear in $h \in X$, there exists a convex neighbourhood U of x_0 such that

$$\|\mathrm{d}f(x;h)\|_Y \leq \|\mathrm{d}f(x)\|_{\mathrm{op}}\|h\|_X \leq K\|h\|_X$$

for all $x \in U$. As *Lipschitz constant K* we can choose the supremum of $x \mapsto \|\mathrm{d}f(x)\|_{\mathrm{op}}$ in U. By the fundamental theorem of calculus we have for $x_1, x_2 \in U$

$$f(x_1) - f(x_2) = \int_0^1 \mathrm{d}f\big(x_2 + t(x_1 - x_2); x_1 - x_2\big)\,\mathrm{d}t.$$

Hence, taking the norm on both sides, the previous equation yields

$$\|f(x_1) - f(x_2)\|_Y \leq \int_0^1 \big\|\mathrm{d}f\big(x_2 + t(x_1 - x_2); x_1 - x_2\big)\big\|_Y\,\mathrm{d}t \leq K\|x_1 - x_2\|_X. \qquad \square$$

Later on we will often encounter spaces of differentiable and smooth functions and thus need an appropriate topology on these space: Let X, Y be a topological vector space such that Y is locally convex and $U \subset X$ open. We can equip the vector space $C^n(U, Y)$ of all n-times continuously differentiable maps between X and Y with the seminorms

$$\|f\|_{i,k,K} = \sup_{x \in K} \|\mathrm{d}^k f(x)\|_{i,\mathrm{op}}$$

for all $f \in C^n(U, Y)$, every compact $K \subset U$ and $0 \leq k \leq n$. These seminorms induce an initial topology on $C^n(U, Y)$ turning it into a locally convex space. This is another example of a compact-open topology or topology of uniform convergence on compacta. Note that, if Y is a Fréchet space and U is σ-compact, $C^n(U, Y)$ becomes a Fréchet space.

3.3.2   The Fréchet derivative

On Banach spaces it is possible to define another directional derivative, the Fréchet derivative. Given Banach spaces X, Y and an open subset $U \subset X$, a map $f : U \to Y$ is called *Fréchet differentiable at $x \in U$* if there exists a bounded linear operator $\mathrm{D}f(x) : X \to Y$, the *Fréchet derivative* of f at x, such that

$$\lim_{\|h\|_X \to 0} \|h\|_X^{-1}\big(f(x+h) - f(x) - \mathrm{D}f(x)h\big) = 0. \tag{3.2}$$

The operator $\mathrm{D}f(x)$ is unique if it exists. In analogy to the directional derivatives that we encountered so far, we also write $\mathrm{D}_h f(x) \doteq \mathrm{D}f(x;h) \doteq \mathrm{D}f(x)h$. We call f *Fréchet differentiable* if the Fréchet derivative exists for all $x \in U$. If the Fréchet derivative is continuous in x, then f is *continuously Fréchet differentiable*.

The Fréchet derivative is closely related to the directional derivative defined above (see also [3]):

Proposition 3.3. *Let X, Y be Banach spaces, $U \subset X$ open and $f : U \to Y$ a map. f is Fréchet differentiable if and only if f is continuously differentiable. In that case the two derivatives agree.*

Proof. "$\Rightarrow$": We can bring (3.2) into agreement with (3.1) by replacing h in (3.2) by th, $t \in \mathbb{R}$, and take the limit $\|th\|_V \to 0$ along the ray of h, *i.e.*, by taking t to zero while keeping h fixed. Moreover, $\mathrm{D}f(x)$ is clearly continuous because it is linear and bounded.

"$\Leftarrow$": As in proposition 3.2, since the derivative $\mathrm{d}f(\cdot)$ is a continuous linear map, there exists a (convex) neighbourhood V of x where it is bounded. Using the fundamental theorem of calculus again, we obtain for any $y \in V$ and sufficiently small $h \in X$

$$\|f(x+h) - f(x) - \mathrm{d}f(y;h)\|_Y \le \sup_{t \in [0,1]} \|\mathrm{d}f(x+th) - \mathrm{d}f(y)\|_{\mathrm{op}} \|h\|_X.$$

In particular this holds for $x = y$ and thus f is Fréchet differentiable at x with $\mathrm{D}f(x) = \mathrm{d}f(x)$. $\qquad\square$

It follows that any statement on continuously differentiable maps also holds for Fréchet differentiable maps.

Fréchet differentiability is a very strong notion of differentiability and many theorems from ordinary calculus can be generalized to the Fréchet derivative but not further to the directional derivative on arbitrary Fréchet spaces. An example is the inverse function theorem for which holds for the Fréchet derivative on Banach spaces but does not hold on general Fréchet spaces. On some Fréchet spaces one has instead the Nash–Moser theorem [116].

3.4   Fixed-point theorems

Let us start this section by stating the most elementary fixed-point theorem, the *Banach fixed-point theorem*:

Theorem 3.4 (Banach fixed-point theorem). *Let $f : X \to X$ be a contraction on a (non-empty) complete metric space X. Then f has a unique fixed-point $x = f(x)$. Furthermore, taking an arbitrary initial value $x_0 \in X$, x is the limit of the sequence (x_n) defined by the iterative procedure $x_{n+1} = f(x_n)$.*

We will not prove this theorem here; the proof is not difficult and can be found in essentially any introductory book on (functional) analysis. Instead we will dissect, specialize, generalize and finally prove various parts of this theorem separately.

3.4.1   Existence and uniqueness

Let us start with a useful lemma:

Lemma 3.5. *Let $\cdots \subset V_k \subset V_{k-1} \subset \cdots \subset V_0$ be a decreasing sequence of sets. Suppose there exists a functional f such that $f : V_k \to V_{k+1}$ for every non-negative $k < n$. Any fixed-point $x = f(x)$ in V_0 is already in V_n.*

Proof. Suppose that $x \in V_0$ but $x \notin V_n$ is a fixed-point. Then there exists a $k < n$ such that $x \in V_k$ and $x \notin V_{k+1}$. Since x is a fixed-point of f, we have that $x = f(x)$, but $f(x) \in V_{k+1}$ by the properties of f.   □

This lemma has serval useful consequences. One example is the following: The limit of a convergent sequence in a complete metric space is not necessarily as regular as all the elements of the sequence; *a priori* the regularity of the limit is only controlled by the topology induced by the metric. However, if the limit is the fixed-point of a smoothing map, the situation is much better.

Corollary 3.6. *Let X, Y be a topological vector spaces and $U \subset X$ open. Further, let $V_k \subset C^k(U, Y)$ for all k such that $V_k \subset V_{k-1}$. Suppose there exists a smoothing functional f such that $f : V_k \to V_{k+1}$ for every non-negative $k < n$. Any fixed-point $x = f(x)$ in C^0 is already in C^n.*   □

If one is interested only in existence of fixed-points but not their uniqueness, then one can perform a straightforward generalization of the 'existence' part of Banach's fixed-point theorem:

Proposition 3.7. *Let (X, d) be a non-empty complete metric space and $f : X \to X$ a map. Assume that there exists a subset $U \subset X$ such that $f : U \to U$ and f is a contraction on U with Lipschitz constant $K \in [0, 1)$, i.e., for all $y, z \in U$*

$$d\big(f(y), f(z)\big) \leq K d(y, z).$$

Then there exists a fixed-point $x = F(x)$ in X.

Proof. Define for an arbitrary $x_0 \in U$ the Picard sequence (x_n) where $x_{n+1} = f(x_n)$. Using the contractivity of f on U, we get

$$d(x_{n+1}, x_n) \leq K d(x_n, x_{n-1}) \leq K^n d(x_1, x_0).$$

One can then easily show that (x_n) is a Cauchy sequence and take the limit $n \to \infty$ in $x_{n+1} = f(x_n)$ to see that there exists a limit $x = f(x)$ in X.   □

This proposition does not guarantee uniqueness of the fixed-point because the mapping is only required to be a contraction on a subset of a complete metric space and the fixed-point is not necessarily contained in this subset. Nevertheless, uniqueness holds if the mapping is of the form assumed in Lem. 3.5. Moreover, if the mapping is smoothing as in Cor. 3.6 then the unique fixed-point is even C^n.

Proposition 3.8. *Let (X, d) be a non-empty complete metric space and (V_k) be a decreasing sequence of sets as in Lem. 3.5 such that $V_0 \subset X$ is closed. Suppose that $f : V_k \to V_{k+1}$ for every non-negative $k < n$ such that f is a contraction on V_n. Then f has a unique fixed-point $x = f(x) \in V_n$.*

Proof. The existence of fixed-points in V_0 that are contained in V_n follows from Lem. 3.5 and Prop. 3.7. Assume now that there exist two distinct fixed-points x, y. Since f is a contraction on V_n, we have

$$d(x, y) = d(f(x), f(y)) \leq K d(x, y),$$

where $K \in [0, 1)$ is the Lipschitz constant of f, and thus arrive at a contradiction. $\square$

3.4.2   A Lipschitz continuity criterion

Next we will see that it is not necessary for a map to be a contraction for it to have fixed points. In fact it is sufficient for the map to satisfy a certain Lipschitz continuity condition:

Lemma 3.9. *Let (X, d) be a non-empty complete metric space. Suppose there exists $K \in \mathbb{R}_+$ such that $f : X \to X$ satisfies*

$$d(f^n(x), f^n(y)) \leq \frac{K^n}{n!} d(x, y)$$

for all $x, y \in X$ and $n \in \mathbb{N}$. Then f has a unique fixed-point.

Proof. Since $n!$ grows faster than K^n, there exists a N such that f^n is a contraction for all $n \geq N$. If we set $V_k = f^k(X)$, we can apply Prop. 3.8 and the thesis follows. $\square$

The special bound assumed in Lem. 3.9 is in fact very natural if f is the intgral functional

$$f : C[a, b] \to C[a, b], \quad f(x)(t) \doteq f_0(t) + \int_a^t k(x)(s) \, \mathrm{d}s, \tag{3.3}$$

where $f_0 \in C[a, b]$ and with integral kernel $k : C[a, b] \to C[a, b]$. Recall that space of continuous functions $C[a, b]$ in the interval $[a, b]$ can be turned into a Banach space by equipping it with with the *uniform norm*

$$\|X\|_{C[a,b]} \doteq \|X\|_\infty \doteq \sup_{t \in [a,b]} |X(t)|,$$

where we will use $\|X\|_{C[a,b]}$ instead of the more common $\|X\|_\infty$ to emphasize the interval over which the supremum is taken.

Proposition 3.10. *Let f be of the form* (3.3) *with k continuously differentiable in $U \subset C[a, b]$ open such that f closes on a closed subset $V \subset U$, i.e., $f(V) \subset V$. Then f has a unique fixed-point in V.*

Proof. We can show the statement using Lem. 3.9 and an inductive procedure. Applying Prop. 3.2, we find that k is locally Lipschitz as a functional on U; denote by $L = \sup_{x \in U} \|\mathrm{d}k\|_{\mathrm{op}}$ its Lipschitz constant. Using the uniform norm on $C[a, t]$, we thus obtain

$$\|f(x) - f(y)\|_{C[a,t]} \leq \int_a^t \|k(x) - k(y)\|_{C[a,t]} \, \mathrm{d}s \leq L(t - a)\|x - y\|_{C[a,b]}.$$

Suppose now that

$$\|f^n(x) - f^n(y)\|_{C[a,t]} \leq \frac{L^n(t - a)^n}{n!} \|x - y\|_{C[a,b]}. \tag{3.4}$$

holds up to n and for arbitrary $t \in [a, b]$. Then,

$$
\begin{aligned}
|f^{n+1}(x)(t) - f^{n+1}(y)(t)| &\leq \int_a^t \|(k \circ f^n)(x) - (k \circ f^n)(y)\|_{C[a,s]} \, \mathrm{d}s \\
&\leq L \int_a^t \|f^n(x) - f^n(y)\|_{C[a,s]} \, \mathrm{d}s \\
&\leq \frac{L^{n+1}}{n!} \int_a^t (s - a)^n \|x - y\|_{C[a,b]} \, \mathrm{d}s \\
&\leq \frac{L^{n+1}(t - a)^{n+1}}{(n+1)!} \|x - y\|_{C[a,b]},
\end{aligned}
$$

which implies that (3.4) holds also for $n + 1$, thus concluding the proof.   □

3.4.3   Closed functionals

The last proposition contains an apparently minor but in fact very strong condition, namely that the functional k closes within the set V. In any application of a fixed-point theorem similar to Banach's theorem, the crucial point to check is usually not that the map is a contraction but that it is closed. In the given case of an integral functional (3.3), however, we can always be assured that there exists an interval $I \subset [a, b]$ on which the functional closes [3]; this interval might be very small.

Proposition 3.11. *Suppose that k is bounded on a set $U \subset C[a, b]$ which also includes a ball V around f_0 defined as $V = \{x \mid \|x - f_0\|_{C[a,b]} < \delta\}$ for some δ. Then there exists $t \in (a, b]$ such that f satisfies $f(U)\!\restriction_{[a,t]} \subset U\!\restriction_{[a,t]}$.*

Proof. Since k is bounded on U, it clearly satisfies

$$\|k(x)\|_{C[a,t]} \leq \|k(x)\|_{C[a,b]} \leq K = \sup_{y \in U} \|k(y)\|_{C[a,b]}$$

for all $x \in U$. Then, taking the norm of (3.3) after subtracting f_0, one obtains

$$\|f(x) - f_0\|_{C[a,t]} \leq (t-a)\|k(x)\|_{C[a,t]} \leq (t-a)K,$$

because $V \subset U$. For any δ we can always find a t such that $(t-a)K < \delta$ and therefore $f(U)\restriction_{[a,t]} \subset V\restriction_{[a,t]}$. The thesis follows because $V \subset U$. □

3.5 Microlocal analysis

3.5.1 Distributions

We will now define three important function spaces and their topological duals, which will be called spaces of distributions.

To conform with standard notation we denote the *space of smooth functions* on an open subset $U \subset \mathbb{R}^n$ by

$$\mathcal{E}(U) \doteq C^\infty(U, \mathbb{C}).$$

As observed in Sect. 3.3.1, it is a Fréchet space with the compact-open topology. The elements of the topological dual $\mathcal{E}'(U)$ are called *compactly supported distributions*.

The vector space of *rapidly decreasing* (or *decaying*) *functions* will be denoted by $\mathcal{S}(U)$. We say that a smooth function $f \in \mathcal{E}(U^n)$ is *rapidly decreasing (decaying)* if

$$\|f\|_{i,n,m} \doteq \sup_{x \in U^n} \left(1 + |x|^n\right)\left\|\mathrm{d}^m f(x)\right\|'_i < \infty$$

for all i, n, m. We equip $\mathcal{S}(U)$ with the topology induced by these seminorms and see that it is a Fréchet space. The topological dual $\mathcal{S}'(U)$ is the space of *tempered distributions* or *Schwartz distributions*.

Another subspace of $\mathcal{E}(U)$ is the space of *test functions*, denoted by

$$\mathcal{D}(U) \doteq C_0^\infty(U, \mathbb{C}).$$

We can equip this space with a topology similar but more complicated than that of $\mathcal{E}(U)$. If $K \subset U$ is compact, we can endow $\mathcal{D}(K) = \mathcal{E}(K)$ with the subspace topology. Then, taking a compact exhaustion $K_1 \subset K_2 \subset \cdots$, $\bigcup_i K_i = U$, the topology on $\mathcal{D}(U)$ is the initial topology defined by the projections $\pi_i : \mathcal{D}(K_i) \to \mathcal{D}(U)$. This topology is not Fréchet unless U is compact and $\mathcal{E}(U)$ is already Fréchet, in which case $\mathcal{D}(U) = \mathcal{E}(U)$. The topological dual $\mathcal{D}'(U)$ is the space of *distributions*.

More generally, we define $\mathcal{F}(U, X)$, with $\mathcal{F} = \mathcal{D}, \mathcal{E}$ or $\mathcal{S}$, as the spaces of functions with values in a locally convex vector space X and by $\mathcal{F}'(U, X)$ the associated distribution spaces. The necessary generalizations to the definitions above are straightforward but note that the resulting function spaces are not Fréchet unless V is already Fréchet. Moreover, it is possible to define Y-valued distributions $\mathcal{F}'(U, X, Y)$, where Y is a locally convex space.

Given a distribution $u \in \mathcal{D}'(U)$, we can *restrict* it to a distribution u_V on any open $V \subset U$ by setting

$$u_V(f) = u(f) \tag{3.5}$$

for every $f \in \mathcal{D}(V)$. A distribution is uniquely determined by its restrictions: If $(U_i)_{i \in \mathbb{N}}$ is an open cover of U and $u_i \in \mathcal{D}'(U_i)$ such that $u_i = u_j$ whenever $U_i \cap U_j \neq \emptyset$, then there exists a unique $u \in \mathcal{D}'(U)$ such that u_i is the restriction of u to U_i for every i.

The *support* $\operatorname{supp} u$ of a distribution $u \in \mathcal{D}'(U)$ is the smallest closed set $V \subset U$ such that the restriction of u to $U \setminus V$ vanishes. More precisely,

$$\operatorname{supp} u = U \setminus \bigcup \{V \subset U \text{ open} \mid u(f) = 0 \; \forall f \in \mathcal{D}(V)\}.$$

It follows that $u(f) = 0$ if $\operatorname{supp} u \cap \operatorname{supp} f = \emptyset$ and that the space $\mathcal{E}'(U)$ is indeed the space of compactly supported distributions.

3.5.2   Distributions on manifolds

The discussion above does not yet encompass the case of distributions on smooth manifolds because manifolds are not vector spaces. However, manifolds are locally homeomorphic to a vector space – Euclidean space.

Let M be a smooth manifold, $E \to M$ a smooth vector bundle and $(U_i)_{i \in \mathbb{N}}$ an open cover of M such that (U_i, φ_i) are coordinate charts and (U_i, ψ_i) local trivializations. We define again the *space of smooth sections*

$$\mathcal{E}(M, E) \doteq \Gamma^\infty(E)$$

as the space of functions $f : M \to E$ such that $\psi_i \circ f \circ \varphi_i^{-1}$ is smooth for each i, *i.e.*, we require

$$\psi_i \circ f \circ \varphi_i^{-1} \in \mathcal{E}\big(\varphi_i(U_i), \psi_i(E_{U_i})\big).$$

A locally convex topology that turns $\mathcal{E}(M, E)$ into a Fréchet spaces is the initial topology induced by the product topology on the right-hand side of the injection

$$\iota : \mathcal{E}(M, E) \to \prod_{i \in \mathbb{N}} \mathcal{E}\big(\varphi_i(U_i), \psi_i(E_{U_i})\big);$$

the topology is independent of the choice of the cover $(U_i)_{i \in \mathbb{N}}$. The topological dual of $\mathcal{E}(M, E)$ is the *space of compactly supported distributional sections* $\mathcal{E}'(M, E)$.

The *space of compactly supported smooth sections*, the *test sections*, is denoted

$$\mathcal{D}(M, E) \doteq \Gamma_0^\infty(E).$$

Analogously to the vector space case we define an initial topology on $\mathcal{D}(M, E)$ induced by that on $\mathcal{E}(K_i, E)$, where (K_i) form a compact exhaustion of M; whence the space of test sections becomes a Fréchet space. The *space of distributional sections* $\mathcal{D}'(M, E)$ is the topological dual of $\mathcal{D}(M, E)$.

The restriction of distribution generalizes to distributions on manifolds in the obvious way: Given a manifold M and an open subset $U \subset M$, every distribution $u \in \mathcal{D}'(M, E)$ can be restricted to a distribution $u_U \in \mathcal{D}'(U, E_U)$ by setting

$$u_U(f) = u(f)$$

for all $f \in \mathscr{D}(U, E_U)$. Also on a manifold a distribution is completely supported by its restrictions.

It does not make sense to define a notion of rapidly decaying sections or tempered distributional sections on manifolds. It is also clear, that the concept of distributions can be further extended to objects such as Fréchet manifolds in very much the same way as above for smooth manifolds.

3.5.3   Nuclearity and the Schwartz kernel theorem

All the function spaces but none of the distribution spaces defined in the previous two sections are Fréchet. However, all the function and distribution spaces (with either the weak or strong topology) are nuclear. For this reason, we have the isomorphisms

$$
\begin{aligned}
\mathscr{D}(U,X) &\simeq \mathscr{D}(U) \widehat{\otimes} X, & \mathscr{D}'(U,\mathbb{K},X) &\simeq \mathscr{D}'(U) \widehat{\otimes} X, \\
\mathscr{E}(U,X) &\simeq \mathscr{E}(U) \widehat{\otimes} X, & \mathscr{E}'(U,\mathbb{K},X) &\simeq \mathscr{E}'(U) \widehat{\otimes} X, \\
\mathscr{S}(\mathbb{R}^m,X) &\simeq \mathscr{E}(\mathbb{R}^m) \widehat{\otimes} X, & \mathscr{S}'(\mathbb{R}^m,\mathbb{K},X) &\simeq \mathscr{S}'(\mathbb{R}^m) \widehat{\otimes} X
\end{aligned}
$$

for a complete locally convex topological vector space X and an open set $U \subset \mathbb{R}^m$. As another consequence we can specialize the abstract kernel theorem (*cf.* Sect. 3.1.5) to these function spaces under which circumstances it is called the *Schwartz kernel theorem*. One finds the following isomorphism (open subsets $U \subset \mathbb{R}^m$ and $V \subset \mathbb{R}^n$):

$$
\begin{aligned}
\mathscr{E}'(U \times V) &\simeq \mathscr{E}'(U) \widehat{\otimes} \mathscr{E}'(V) \simeq L\big(\mathscr{E}(U), \mathscr{E}'(V)\big), \\
\mathscr{D}'(U \times V) &\simeq \mathscr{D}'(U) \widehat{\otimes} \mathscr{D}'(V) \simeq L\big(\mathscr{D}(U), \mathscr{D}'(V)\big), \\
\mathscr{S}'(\mathbb{R}^{m+n}) &\simeq \mathscr{S}'(\mathbb{R}^m) \widehat{\otimes} \mathscr{S}'(\mathbb{R}^n) \simeq L\big(\mathscr{E}(\mathbb{R}^m), \mathscr{E}'(\mathbb{R}^n)\big),
\end{aligned}
$$

where $L(X,Y)$ denotes the space of continuous linear maps between topological vector spaces X and Y with the topology of uniform convergence. Analogous isomorphisms (at least for $\mathscr{E}$ and $\mathscr{D}$) exist for both sets of isomorphisms also for functions and distributions on manifolds.

As a consequence of these isomorphisms, there exists for every distribution $K \in \mathscr{D}'(U \times V)$ a unique linear operator $\mathrm{K} : \mathscr{D}(U) \to \mathscr{D}'(V)$ and, conversely, to every linear operator K a unique distribution. Let $f \in \mathscr{D}(U)$ and $g \in \mathscr{D}(V)$ be test functions. Formally we can write

$$
K(f \otimes g) = \int_{U \times V} K(x,y) f(x) g(x) \, \mathrm{d}^m x \, \mathrm{d}^n y
$$

for the distribution with *distributional kernel* $K(x,y)$ and

$$
(\mathrm{K}f)(y) = \int_U K(x,y) f(x) \, \mathrm{d}^m x, \quad ({}^t\mathrm{K}g)(x) = \int_V K(x,y) g(y) \, \mathrm{d}^n y,
$$

for the associated operator and its transpose.

The operator K is called *semiregular* if it continuously[5] maps $\mathscr{D}(U)$ into $\mathscr{E}(V)$ and, analogously, the transpose ${}^t\mathrm{K}$ is called semiregular if it continuously maps $\mathscr{D}(V)$ into

---

[5]Continuity is meant with respect to the usual topology of $\mathscr{E}(V)$ and not the subspace topology of $\mathscr{D}'(V)$.

$\mathcal{E}(U)$. In the case that ^{t}K is semiregular, we can uniquely extend K to an operator acting on compactly supported distributions $\mathcal{E}'(U)$ by duality:

$$(Ku)(g) = ({}^t Kg)(u)$$

for all $u \in \mathcal{E}(U)$ and $g \in \mathcal{D}'(V)$.

If both K and ^{t}K are semiregular, we say that K *regular*. Moreover, it is called *properly supported* if the projections from $\operatorname{supp} K \subset U \times V$ onto each factor are proper maps. A properly supported operator K maps $\mathcal{D}(U)$ to $\mathcal{E}'(V)$ and can therefore be extended to an operator $\mathcal{D}(U) \to \mathcal{E}'(V)$. Since linear differential operators are properly supported and regular, they can be uniquely extended to distributions and they can also be composed.

3.5.4   Fourier transformation and convolution

Let us denote by $L^p(\mathbb{R}^n)$ the L^p *spaces* of functions on $\mathbb{R}^n$ with values in $\mathbb{C}$. That is, $L^p(\mathbb{R}^n)$ is the space of functions for which the Lebesgue integrals

$$\|f\|_p \doteq \left(\int_{\mathbb{R}^n} |f(x)|^p \, \mathrm{d}^n x \right)^{1/p}$$

exist and where we identify functions which are equal almost everywhere so that the L^p spaces become Banach spaces. L^1 functions are called *Lebesgue integrable*, while L^2 are called *square-integrable*.

On the space of Lebesgue integrable functions $L^1(\mathbb{R}^n)$, the *Fourier transform* is defined as the automorphism

$$\mathcal{F} : f(x) \mapsto \mathcal{F}(f)(\xi) \doteq (2\pi)^{-n} \int_{\mathbb{R}^n} f(x) \mathrm{e}^{-\mathrm{i} x \cdot \xi} \, \mathrm{d}^n x,$$

where $\cdot$ denotes the Euclidean dot product. The Fourier transform satisfies

$$\mathcal{F}^2(f)(x) = (2\pi)^n f(-x)$$

and the *inverse Fourier transform* is therefore given by

$$\mathcal{F}^{-1}(f)(x) \doteq \int_{\mathbb{R}^n} f(\xi) \mathrm{e}^{\mathrm{i} x \cdot \xi} \, \mathrm{d}^n \xi.$$

When no confusion can arise, we usually denote the Fourier transform of a function f by $\widehat{f}$ instead of $\mathcal{F}(f)$.

By the Riemann–Lebesgue lemma, it is clear that $\widehat{f}(\xi) \to 0$ as $|\xi| \to \infty$. In fact, the Fourier transform is a linear isomorphism from the subspace of rapidly decaying functions $\mathcal{S}(\mathbb{R}^n)$ into itself. Since the space of rapidly decreasing functions is stable under differentiation and multiplication by polynomials, one finds for $f \in \mathcal{S}(\mathbb{R}^n)$

$$\mathcal{F}(\partial_\mu f)(\xi) = \xi_\mu \widehat{f}(\xi) \quad \text{and} \quad \mathcal{F}(x_\mu f)(\xi) = \partial_\mu \widehat{f}(\xi). \tag{3.6}$$

Moreover, given also $g \in \mathcal{S}(\mathbb{R}^n)$, the *Plancherel–Parseval identities* are

$$\int_{\mathbb{R}^n} f(x)\overline{g}(x)\,\mathrm{d}^n x = (2\pi)^{-n} \int_{\mathbb{R}^n} \widehat{f}(\xi)\overline{\widehat{g}}(\xi)\,\mathrm{d}^n \xi, \qquad (3.7)$$

$$\int_{\mathbb{R}^n} |f(x)|^2\,\mathrm{d}^n x = (2\pi)^{-n} \int_{\mathbb{R}^n} |\widehat{f}(\xi)|^2\,\mathrm{d}^n \xi.$$

As a consequence, the Fourier transform can be extended to an isomorphism of $L^2(\mathbb{R}^n)$ into itself.

The Plancherel–Parseval formula (3.7) guides us to extend the Fourier transformation $\mathcal{F}$ further to the space of tempered distributions $\mathcal{S}'(\mathbb{R}^n)$ by

$$\langle \widehat{u}, f \rangle \doteq \langle u, \widehat{f} \rangle$$

for all $u \in \mathcal{S}'(\mathbb{R}^n)$ and $f \in \mathcal{S}(\mathbb{R}^n)$, *i.e.*, it is the transpose of the Fourier transformation on rapidly decreasing functions. It follows that $\mathcal{F}$ is a linear isomorphism from $\mathcal{S}'(\mathbb{R}^n)$ (with the weak topology) into itself and analogues of the relations (3.6) hold also for tempered distributions $u \in \mathcal{S}'(\mathbb{R}^n)$:

$$\mathcal{F}(\partial_\mu u)(\xi) = \xi_\mu \widehat{u}(\xi) \quad \text{and} \quad \mathcal{F}(x_\mu u)(\xi) = -\partial_\mu \widehat{u}(\xi).$$

If we restrict to the space of compactly supported distributions, the Fourier transform of $u \in \mathcal{E}'(\mathbb{R}^n)$ is equivalently given as the smooth function

$$\widehat{u}(\xi) = \langle u, f : x \mapsto \mathrm{e}^{-\mathrm{i}x\cdot\xi} \rangle.$$

The Fourier transform $\widehat{u}$ can be directly extended to $\mathbb{C}^n$ as an entire analytic function.

The *convolution* of two Lebesgue integrable functions $f, g \in L^1(\mathbb{R}^n)$ is defined as

$$(f * g)(x) \doteq \int_{\mathbb{R}^n} f(y)g(x - y)\,\mathrm{d}^n y.$$

The product thus defined is dual to the usual product with respect to Fourier transformation. To wit, the identities

$$\mathcal{F}(f * g) = \widehat{f}\,\widehat{g} \quad \text{and} \quad \mathcal{F}(f\,g) = (2\pi)^{-n}(\widehat{f} * \widehat{g})$$

hold and are the result of the *convolution theorem*. Note that for distributions $u \in \mathcal{S}'(\mathbb{R}^n)$ and $v \in \mathcal{E}'(\mathbb{R}^n)$, the convolution $u * v$ is a well-defined tempered distribution and its Fourier transform satisfies $\mathcal{F}(u * v) = \widehat{u}\widehat{v}$ as in the convolution theorem. If also the product uv is well-defined as a (tempered) distribution, *cf.* Sect. 3.5.7), then other statement of the convolution theorem holds and $\mathcal{F}(uv) = (2\pi)^{-n}(\widehat{u} * \widehat{v})$

3.5.5   Singularities and the wavefront set

Every locally Lebesgue integrable function $u \in L^1_{\mathrm{loc}}(\mathbb{R}^n)$ can be identified with a distribution in $\mathcal{D}'(\mathbb{R}^n)$, denoted by the same symbol, via

$$u(f) = \langle u, f \rangle = \int_{\mathbb{R}^n} u(x)f(x)\,\mathrm{d}^n x$$

for all $f \in \mathcal{D}(\mathbb{R}^n)$. We say that a distribution $u \in \mathcal{D}'(\mathbb{R}^n)$ is *smooth*, if it is induced from a smooth function via this duality pairing. More specifically, every smooth function corresponds to a distribution in this way and in fact $\mathcal{D}(\mathbb{R}^n)$ is isomorphic to a dense subset of $\mathcal{D}'(\mathbb{R}^n)$. Therefore this pairing uniquely extends to a pairing between $\mathcal{D}(\mathbb{R}^n)$ and $\mathcal{D}'(\mathbb{R}^n)$. Using the Plancherel–Parseval identity (3.7), it can be written explicitly as

$$u(f) = (2\pi)^{-n} \int_{\mathbb{R}^n} \widehat{\chi u}(\xi) \widehat{f}(-\xi) \, \mathrm{d}^n \xi, \qquad (3.8)$$

where $u \in \mathcal{D}'(\mathbb{R}^n)$ and $\chi \in \mathcal{D}(\mathbb{R})$ such that $\chi = 1$ on a compact neighbourhood of the support of $f \in \mathcal{D}(\mathbb{R}^n)$. This pairing may be considered the motivation of the wavefront set to be defined below.

The *singular support* $\mathrm{sing\,supp}\, u$ of a distribution $u \in \mathcal{D}'(\mathbb{R}^n)$ is then defined as the complement of the union of all open sets on which u is smooth in the sense of the pairing above. In other words, it is the smallest closed subset $U \subset \mathbb{R}^n$ such that $u_{\mathbb{R}^n \setminus U} \in \mathscr{E}(\mathbb{R}^n \setminus U)$.

The Fourier transform, introduced in the previous section, can be used to give a condition on the smoothness of a compactly supported distribution $u \in \mathscr{E}'(\mathbb{R}^n)$. Namely, u is smooth if and only if for each $n \in \mathbb{N}_0$ there exists a constant C_n such that

$$|\widehat{u}(\xi)| \le C_n (1 + |\xi|)^{-n}$$

for all $\xi \in \mathbb{R}^n$.

Checking this condition for certain ξ, a *regular direction* of a compactly supported distribution $u \in \mathscr{E}'(\mathbb{R}^n)$ is a vector $\xi \in \mathbb{R}^n \setminus \{0\}$ such that there exists an open conical[6] neighbourhood Γ of ξ and such that

$$\sup_{\zeta \in \Gamma} (1 + |\zeta|)^n |\widehat{u}(\zeta)| < \infty$$

for all $n \in \mathbb{N}_0$. Conversely, a ξ is called a *singular direction* of u if it is not a regular direction. The (closed) *set of all singular directions* of u is

$$\Sigma(u) \doteq \{\xi \in \mathbb{R}^n \setminus \{0\} \mid \xi \text{ is not a regular direction of } u\},$$

i.e., the complement of all regular directions.

We can localize the notion of singular directions and say that ξ is a singular direction of $u \in \mathcal{D}'(U)$ at $x \in U$, where $U \subset \mathbb{R}^n$ is open, if there exists a $n \in \mathbb{N}_0$ such that

$$\sup_{\zeta \in \Gamma} (1 + |\zeta|)^n |\mathcal{F}(\chi u)(\zeta)|$$

is not bounded for all $\chi \in \mathcal{D}(U)$ localized at x (*i.e.*, $\chi(x) \ne 0$). That is, the *set of singular directions at x* is the closed set

$$\Sigma_x(u) \doteq \bigcap_\chi \Sigma(\chi u),$$

---

[6] A *cone* in $\mathbb{R}^n$ is a subset $\Gamma \subset \mathbb{R}^n$ such that $\lambda \Gamma = \Gamma$ for all $\lambda > 0$.

where the intersection is over all $\chi \in \mathscr{D}(U)$ such that $\chi(x) \neq 0$.

This leads to the definition of the *wavefront set* as the set of the singular directions at all points:

$$\mathrm{WF}(u) \doteq \big\{ (x;\xi) \in U \times (\mathbb{R}^n \setminus \{0\}) \,\big|\, \xi \in \Sigma_x(u) \big\}.$$

Thus the wavefront set is a refinement of the notion of singular support. Moreover, it can be used as a practical tool for calculating the singular support because $\mathrm{sing\,supp}\,u$ is the projection of $\mathrm{WF}(u)$ onto the first component. The wavefront set has the following properties:

(a) $\mathrm{WF}(\chi u) \subset \mathrm{WF}(u)$,

(b) $\mathrm{WF}(u + v) \subset \mathrm{WF}(u) \cup \mathrm{WF}(v)$,

(c) $\mathrm{WF}(\mathrm{P}u) \subset \mathrm{WF}(u)$

for all distributions $u, v \in \mathscr{D}'(U)$, localizing functions $\chi \in \mathscr{D}(U)$ and linear differential operators P (with smooth coefficients).

3.5.6 Wavefront set in cones

Let $U \subset \mathbb{R}^n$ be open and $\Gamma \subset U \times (\mathbb{R}^n \setminus \{0\})$ a closed cone, where we have extended the definition of a cone to sets for which the projection to the second component at each point is a cone. We define distributions with wavefront set contained in the cone Γ as

$$\mathscr{D}'_\Gamma(U) \doteq \big\{ u \in \mathscr{D}'(U) \,\big|\, \mathrm{WF}(u) \subset \Gamma \big\},$$

which is not empty for any cone Γ. The *normal topology*[7] turns $\mathscr{D}'_\Gamma(U)$ into a complete nuclear space [54]. It is induced by the seminorms

$$\|u\|_B = \sup_{f \in B} |u(f)| \quad \text{and} \quad \|u\|_{n,V,\chi} = \sup_{\xi \in V} (1 + |\xi|)^n |\mathscr{F}(\chi u)(\xi)|,$$

for all bounded sets $B \subset \mathscr{D}(U)$, $n \in \mathbb{N}_0$, localizing functions $\chi \in \mathscr{D}(U)$ and closed cones $V \subset \mathbb{R}^n \setminus \{0\}$ such that $\mathrm{supp}(\chi) \times V \subset \Gamma$.

Given a closed cone Γ as above, define the open cone

$$\Lambda = (\Gamma')^c \doteq \big\{ (x;\xi) \in U \times (\mathbb{R}^n \setminus \{0\}) \,\big|\, (x;-\xi) \notin \Gamma \big\}$$

as the complement of the reflection of Γ and

$$\mathscr{E}'_\Lambda(\mathbb{R}^n) \doteq \big\{ v \in \mathscr{E}'(\mathbb{R}^n) \,\big|\, \mathrm{WF}(v) \subset \Lambda \big\}$$

as the space of compactly supported distributions with wavefront set contained in Λ. Then one can find an analogue to the pairing (3.8) for all $u \in \mathscr{D}'_\Gamma(\mathbb{R}^n)$ and $v \in \mathscr{E}'_\Lambda(\mathbb{R}^n)$ given by [54]

$$\langle u, v \rangle \doteq (2\pi)^{-n} \int_{\mathbb{R}^n} \widehat{\chi u}(\xi) \widehat{v}(-\xi) \, \mathrm{d}^n \xi,$$

---

[7]The normal topology [54] is finer than the often emplyed *Hörmander topology* for these spaces. Nuclearity also holds for the Hörmander topology but not completeness.

for any $\chi \in \mathscr{D}(\mathbb{R})$ such that $\chi = 1$ on a compact neighbourhood of supp v. $\mathscr{E}'_\Lambda(\mathbb{R}^n)$ thus becomes the topological dual of $\mathscr{D}'_\Gamma(\mathbb{R}^n)$ (with the normal topology) and (equipped with the strong topology) it is also nuclear but not complete unless Λ is also closed [54].

3.5.7   Pullback of distributions

Let U, V be open subsets of $\mathbb{R}^n$ and $\iota : U \to V$ a diffeomorphism. The pullback $\iota^* u$ of a distribution $u \in \mathscr{D}'(V)$ is (uniquely) defined for every $f \in \mathscr{D}(U)$ as the transpose of the pushforward (up to the Jacobian determinant)

$$\langle \iota^* u, f \rangle = \langle u, \iota_* f \, |\det d\iota| \rangle$$

or, equivalently, as the continuous extension of the pullback on smooth function. Consequently, for any closed cone $\Gamma \subset V \times (\mathbb{R}^n \setminus \{0\})$, one obtains

$$\iota^* \mathscr{D}'_\Gamma(V) = \mathscr{D}'_{\iota^* \Gamma}(U), \quad \iota^* \Gamma \doteq \left\{ (x; T_x^* \iota(\xi)) \mid (\iota(x); \xi) \in \Gamma \right\} \tag{3.9}$$

and hence $\mathrm{WF}(\iota^* u) = \iota^* \mathrm{WF}(u)$.

Trying to generalize this result to cases where $\iota : U \to V$ is not a diffeomorphism but an embedding of an open subset of $\mathbb{R}^n$ into an open subset of $\mathbb{R}^m$ can fail if there are $(x; \xi)$ such that $T_x^* \iota(\xi) = 0$. It follows that a distribution $u \in \mathscr{D}'(V)$ can only be pulled back to a distribution $\iota^* u$ if $\mathrm{WF}(u) \cap N = \emptyset$, where

$$N = \left\{ (\iota(x); \xi) \in V \times \mathbb{R}^m \mid x \in U, T_x^* \iota(\xi) = 0 \right\}$$

is the set of conormals of ι.

Given two distributions $u \in \mathscr{D}'(U)$ and $v \in \mathscr{D}'(V)$, where $U \subset \mathbb{R}^n$ and $V \subset \mathbb{R}^m$ are open, the tensor product

$$u \otimes v : f \otimes h \mapsto u(f)v(h)$$

is a distribution in $\mathscr{D}'(U \times V) \simeq \mathscr{D}'(U) \widehat{\otimes} \mathscr{D}'(V)$ via Schwartz's kernel theorem. One can show that its wavefront set satisfies

$$\mathrm{WF}(u \otimes v) \subset \big(\mathrm{WF}(u) \times \mathrm{WF}(v)\big) \cup \big((\mathrm{supp}\, u \times \{0\}) \times \mathrm{WF}(v)\big)$$
$$\cup \big(\mathrm{WF}(u) \times (\mathrm{supp}\, v \times \{0\})\big).$$

It is possible to pullback the tensor product $u \otimes v$ of two distributions over the same space (*i.e.*, $U = V$) with the diagonal map

$$\Delta : U \times U \to U, (x, x) \mapsto x$$

if $\mathrm{WF}(u \otimes v) \cap N_\Delta = \emptyset$, where N_Δ is the set of conormals with respect to the map Δ, which gives the (unique) product uv of the two distributions. This requirement of the wavefront set implies that it is possible to multiply two distributions if and only if

$$(x, \xi) \in \mathrm{WF}(u) \implies (x, -\xi) \notin \mathrm{WF}(v) \tag{3.10}$$

and then wavefront set of the product is bounded by

$$\mathrm{WF}(uv) \subset \left\{ (x;\xi+\zeta) \,\middle|\, (x;\xi) \in \mathrm{WF}(u), (x;\zeta) \in \mathrm{WF}(v) \right\}$$
$$\cup \mathrm{WF}(u) \cup \mathrm{WF}(v). \tag{3.11}$$

Note that for u,v that do not satisfy (3.10), the singular directions would add up to zero in the first term on the right-hand side of (3.11).

3.5.8 Wavefront set of distributional sections

The wavefront set can be extended to distributions on vector-valued functions component-wise, *i.e.*, using $\mathscr{D}'(U,\mathbb{K}^m) \simeq \mathscr{D}'(U) \otimes \mathbb{K}^m \simeq \mathscr{D}'(U)^{\oplus m}$. Namely, one defines for $u \in \mathscr{D}'(U,\mathbb{K}^m)$

$$\mathrm{WF}(u) \doteq \bigcup_{i=1}^{m} \mathrm{WF}(u_i),$$

where $u_i \in \mathscr{D}'(U)$ are the components of u. This definition is invariant under a change of basis because such a change only implies a multiplication of (u_i) by a matrix with smooth components.

Moreover, the wavefront set being a local concept, it generalizes to manifolds and distributional sections in a coordinate neighbourhood via local trivializations. However, to be meaningful, it needs to transform covariantly under diffeomorphisms.

Let $(U_i)_{i \in \mathbb{N}}$ be an open cover of a smooth n-manifold M such that (U_i, φ_i) are coordinate charts and (U_i, ψ_i) are local trivializations of the vector bundle $E \to M$. Given a distribution $u \in \mathscr{D}'(M,E)$ with restrictions u_i to U_i, the wavefront set for every u_i given by

$$\mathrm{WF}(u_i) \doteq \left\{ (x; T_x^*\varphi(\xi)) \in U_i \times (\mathbb{R}^n \setminus \{0\}) \,\middle|\, (\varphi(x); \xi) \in \mathrm{WF}(\psi_i \circ u_i \circ \varphi_i^{-1}) \right\}.$$

and transforms as a conical subset of the cotangent bundles T^*U_i as seen by (3.9). In particular, $\mathrm{WF}(u_i) \cap T^*(U_i \cap U_j) = \mathrm{WF}(u_j) \cap T^*(U_i \cap U_j)$ for all i,j.

The *wavefront set of distributional sections u* is then defined as the union of all $\mathrm{WF}(u_i)$. In other words, it is the set of points

$$(x;\xi) \in \dot{T}^*M \doteq T^*M \setminus \{(y;0) \in T^*M\},$$

the cotangent bundle with the zero section removed, such that $(x;\xi) \in \mathrm{WF}(u_U)$, where U is a coordinate and trivialization neighbourhood of E.

3.5.9 Some distributions and their wavefront set

For any $f \in \mathscr{E}(\mathbb{R})$, *Dirac's δ-distribution* is

$$\delta(f) = f(0)$$

and it follows that δ has support only at the origin. There it does not decay in any direction because $\widehat{\delta} = 1$ so that $\mathrm{WF}(\delta) = \{0\} \times (\mathbb{R} \setminus \{0\})$. Consequently powers of the δ-distribution cannot be defined.

The Dirac δ-distribution can be decomposed into two distributions

$$\delta_\pm(f) \doteq \lim_{\varepsilon \to 0^+} \int_{\mathbb{R}} \frac{f(x)}{x \pm i\varepsilon} \, dx,$$

for all $f \in \mathcal{S}(\mathbb{R})$, such that $-2\pi i\delta = \delta_+ + \delta_-$, where $\mathrm{WF}(\delta_\pm) = \{0\} \times \mathbb{R}_\pm$. Now, powers of either $\delta_\pm$ are well-defined but the distribution $\delta_+\delta_-$ does not exist.

The wavefront set of δ_+ (and analogously that of δ_-) can be calculated as follows: Using the residue theorem, the Fourier transform of $1/(x + i\varepsilon)$ for $\varepsilon > 0$ is[8]

$$\int_{\mathbb{R}} \frac{e^{-ix\xi}}{x + i\varepsilon} \, dx = -2\pi i\theta(\xi)\,e^{-\xi\varepsilon}.$$

Taking the limit $\varepsilon \to 0^+$, this gives $\widehat{\delta}_+(\xi) = -2\pi i\theta(\xi)$. Then, applying the convolution theorem, one obtains the Fourier transform of $\chi\delta_+$ for all $\chi \in \mathcal{D}(\mathbb{R})$ as

$$\mathcal{F}(\chi\delta_+) = \frac{1}{2\pi}(\widehat{\chi} * \widehat{\delta}_+) = -i\int_{-\infty}^{\xi} \widehat{f}(k) \, dk.$$

Since this decays rapidly as $\xi \to -\infty$ and does not decay as $\xi \to \infty$, we get the expected wavefront set.

Related to the diagonal map $\Delta : (x, x) \mapsto x$, we can define for all $f \in \mathcal{D}(\mathbb{R}^2)$ a *diagonal distribution*

$$\Delta(f) \doteq \int_{\mathbb{R}} f(x, x) \, dx.$$

It is clear that the wavefront set of Δ is

$$\mathrm{WF}(\Delta) = N_\Delta = \{(x, x; \xi, -\xi) \in \mathbb{R}^4 \setminus \{0\}\}.$$

Given instead two functions $f_1, f_2 \in \mathcal{D}(\mathbb{R})$, we can write

$$\Delta(f_1 \otimes f_2) = \int_{\mathbb{R}} (f_1 * \delta)(x) f_2(x) \, dx$$

Splitting the δ-distribution into its positive and negative frequency components as above, we can therefore define

$$\Delta_\pm(f) \doteq \lim_{\varepsilon \to 0^+} \int_{\mathbb{R}} \frac{f(x, y)}{y - x \pm i\varepsilon} \, dx \, dy,$$

which in the case $f = f_1 \otimes f_2$ can be written as

$$\Delta_\pm(f_1 \otimes f_2) = \int_{\mathbb{R}} (f_1 * \delta_\pm)(x) f_2(x) \, dx.$$

It is not a difficult exercise to show that [5, Exmpl. 1.4]

$$\mathrm{WF}(\Delta_\pm) = \{(x, x; \xi, -\xi) \in \mathbb{R}^4 \setminus \{0\} \mid \pm\xi > 0\}.$$

Moreover, using the Plancherel–Parseval identities it is possible to show that $\Delta_\pm$ is well-defined for all $f_1, f_2 \in L^2(\mathbb{R})$.

---

[8] θ denotes the Heaviside step-function.

3.5.10   Propagation of singularities

In Sect. 3.5.5 we already noticed that $\mathrm{WF}(\mathrm{P}u) \subset \mathrm{WF}(u)$. That is, knowing the wavefront set of the distribution u, we can deduce information about the wavefront set of $\mathrm{P}u$, where P is a differential operator. The theorem on the *propagation of singularities* gives us information in the opposite direction. Namely, $\mathrm{WF}(\mathrm{P}u)$ and the form of P, tell us a lot about $\mathrm{WF}(u)$.

Let $\mathrm{P} : \mathscr{E}(M,E) \to \mathscr{E}(M,E)$ be a differential operator acting on sections of a vector bundle $E \to M$. Its *characteristic set* is the cone

$$\mathrm{char}\,\mathrm{P} = \left\{ (x;\xi) \in \dot{T}^{*}M \;\middle|\; \det \sigma_{\mathrm{P}}(x,\xi) = 0 \right\}$$

on which the principal symbol σ_{P} of P cannot be inverted.[9] An integral curve of σ_{P} in $\mathrm{char}\,\mathrm{P}$ is called a *bicharacteristic strip*, its projection onto M a *bicharacteristic*.

Theorem 3.12 (Propagation of singularities). *Suppose that* P *is a differential operator with real homogeneous principal symbol such that no complete bicharacterstic stays in a compact set of* M *(i.e.,* P *is of real principal type) and let* $u, f \in \mathscr{D}'(M,E)$ *such that* $\mathrm{P}u = f$. *Then*

$$\mathrm{WF}(u) \subset \mathrm{char}\,\mathrm{P} \cup \mathrm{WF}(f)$$

and, if $(x;\xi) \in \mathrm{WF}(u) \setminus \mathrm{WF}(f)$, *it follows that* $(x';\xi') \in \mathrm{WF}(u)$ *for all* $(x';\xi')$ *on the bicharacteristic strip passing through* $(x;\xi)$.

3.6   Wave equations

Both classical and quantum fields usually satisfy an equation of motion given by a *wave equation*

$$\mathrm{P}u = f, \tag{3.12}$$

where P is a normally hyperbolic differential operator, u is the field and f and an external source. On globally hyperbolic manifolds the wave equation can be solved, *i.e.*, the Cauchy problem for (3.12) is well-posed.

3.6.1   Retarded and advanced propagators

Let (M,g) be a spacetime and $\mathrm{P} : \mathscr{E}(M,E) \to \mathscr{E}(M,E)$ a differential operator on sections of a vector bundle $E \to M$. A linear operator $\mathrm{G}_{\vee} : \mathscr{D}(M,E) \to \mathscr{E}(M,E)$ such that for all $f \in \mathscr{D}(M,E)$

$$\mathrm{P}\mathrm{G}_{\vee}f = f \quad \text{and} \quad \mathrm{G}_{\vee}\mathrm{P}f = f,$$

i.e., $\mathrm{G}_{\vee}$ is a left- and right-inverse of P, and

$$\mathrm{supp}(\mathrm{G}_{\vee}f) \subset J^{+}(\mathrm{supp}\,f)$$

---

[9]If the principal symbol of P is invertible, we say that P is *elliptic*. Example of elliptic operators are the Laplace operator and the Cauchy–Riemann operator.

is called a *retarded propagator* or *retarded Green's operator* for P. Similarly, a linear operator $G_\wedge$, which is a two-sided inverse of P and satisfies

$$\operatorname{supp}(G_\wedge f) \subset J^-(\operatorname{supp} f)$$

for all test sections f, is called a *advanced propagator* or *advanced Green's operator*.[10] We say that P is *Green-hyperbolic* if it admits unique retarded and advanced propagators when restricted to a globally hyperbolic region.

Given a linear differential operator Q such that $P \circ Q = Q \circ P$, *i.e.*, Q commutes with P, then it also commutes with the propagators of P. That is, one finds

$$G_\vee Q f = Q G_\vee f \quad \text{and} \quad G_\wedge Q f = Q G_\wedge f$$

for all $f \in \mathscr{E}(M, E)$.

If P is Green-hyperbolic, then the transpose operator ${}^t P$ on sections of the dual bundle E^* is also Green-hyperbolic; we denote its propagators by $G_\vee^t$ and $G_\wedge^t$. They are closely related to the propagators of P and one finds

$$G_\vee = {}^t(G_\wedge^t) \quad \text{and} \quad G_\wedge = {}^t(G_\vee^t).$$

Since the propagators are regular, they can be uniquely extended to operators $\mathscr{E}'(M, E) \to \mathscr{D}'(M, E)$. Although the propagators are not properly supported, they can also be defined for some non-compactly supported sections. The geometry of (M, g) enables us to define further types of 'compact' support: We say that a (distributional) section u is future or past compact if there exists a Cauchy surface Σ such that

$$\operatorname{supp} u \subset J^+(\Sigma) \quad \text{or} \quad \operatorname{supp} u \subset J^-(\Sigma),$$

respectively. Denote by the subscripts 'fc' and 'pc' the subsets of (distributional) sections of future and past compact support. Via the transpose propagators $G_\vee^t, G_\wedge^t$, we can then uniquely extend the retarded propagator to $\mathscr{D}'_{\mathrm{fc}}(M, E) \to \mathscr{D}'(M, E)$ and the advanced propagator to $\mathscr{D}'_{\mathrm{pc}}(M, E) \to \mathscr{D}'(M, E)$.

Let E be endowed with a bundle metric $(\cdot, \cdot)$. The *formal adjoint* P^* of P with respect to $(\cdot, \cdot)$ is given by

$$\int_M (Pf, h)\,\mu_g = \int_M (f, P^* h)\,\mu_g$$

for all $f, h \in \mathscr{E}(M, E)$ such that $\operatorname{supp} f \cap \operatorname{supp} h$ is compact. If $P^* = P$, the operator is called *formally self-adjoint*. In that case, it follows from the last paragraph that

$$\int_M (G_\vee f, h)\,\mu_g = \int_M (f, G_\wedge h)\,\mu_g.$$

As indicated above, wave operators on globally hyperbolic manifolds play an important role and, in fact, they are particularly well-behaved [24, 103]:

---

[10]Note that our definition of the support of the retarded and advanced propagators is exactly opposite to that in [23, 24] and also [72].

Theorem 3.13. *Any normally hyperbolic operator* P *on a globally hyperbolic manifold admits unique retarded* $G_\vee$ *and advanced propagators* $G_\wedge$.

It is not difficult to extend this result to pre-normally hyperbolic operators on globally hyperbolic spacetimes. Namely, given pre-normally hyperbolic operators P and Q such that $P \circ Q$ is normally hyperbolic, P possesses unique retarded and advanced propagators

$$\widetilde{G}_\vee = Q \circ G_\vee \quad \text{and} \quad \widetilde{G}_\wedge = Q \circ G_\wedge,$$

where $G_\vee, G_\wedge$ are the propagators for the composite operator $P \circ Q$.

3.6.2 Causal propagator

The *causal propagator* is defined as the difference of the retarded and advanced propagator

$$G \doteq G_\vee - G_\wedge.$$

From the support properties of the retarded and advanced propagator it is clear that $\mathrm{supp}(Gf) = J(\mathrm{supp}\, f)$ for all $f \in \mathscr{D}(M, E)$. In Sect. 5.3.2 we will see that the causal propagator, or rather the associated distribution via Schwartz's kernel theorem, may also be called the *commutator distribution* or *Pauli–Jordan distribution*.

By the regularity of the retarded and advanced propagators, it is clear that G extends to an operator $\mathscr{E}'(M, E) \to \mathscr{D}'(M, E)$. Noting the support property of G, this statement can be strengthened to extend the causal propagator to $\mathscr{D}'_{tc}(M, E) \to \mathscr{D}'(M, E)$. Here we have denoted by a subscript 'tc' the space of (distributional) sections of timelike compact support, *i.e.*, the sections u such that

$$\mathrm{supp}\, u \subset J^+(\Sigma_1) \cap J^-(\Sigma_2)$$

for two Cauchy surfaces Σ_1, Σ_2.

Every smooth and spacelike compact solution of the homogeneous differential equation $Pu = 0$ propagating on a globally hyperbolic spacetime (M, g) with Green-hyperbolic operator P can be obtained by applying G to a test section f. In fact, if we denote by $\mathscr{E}_{sc}(M, E)$ the smooth sections of E with spacelike compact support, then we find the exact sequence

$$\{0\} \longrightarrow \mathscr{D}(M, E) \xrightarrow{\ P\ } \mathscr{D}(M, E) \xrightarrow{\ G\ } \mathscr{E}_{sc}(M, E) \xrightarrow{\ P\ } \mathscr{E}_{sc}(M, E).$$

This sequence also entails the fact that the kernel of G is given by $P\mathscr{D}(M, E)$. In other words, $f - f' = Ph$ for some $f, f', h \in \mathscr{D}(M, E)$ implies that $Gf = Gf'$.

Closely related to the existence of a causal propagator is the question whether the *Cauchy problem* is well-posed. The Cauchy problem for the wave equation $Pu = 0$ on a globally hyperbolic manifold (M, g) is the following:

Given a Cauchy surface $\iota : \Sigma \to M$ with normal vector field n, does there exist a unique section $u \in \mathscr{E}(M, E)$ such that

$$\begin{cases} Pu = 0, \\ \iota^* u = u_0, \\ \iota^* \nabla_n u = u_1 \end{cases}$$

and the solution u depends continuously on the data $u_0, u_1 \in \mathscr{E}(\Sigma, \iota^* E)$.

This question can be answered in the positive for normally hyperbolic operators P on globally hyperbolic spacetimes. With the appropriate modifications, the Cauchy problem can also be formulated for pre-normally hyperbolic operators. Also in that case Cauchy problem is well-posed [220]. For general Green-hyperbolic operators the Cauchy problem is more complicated and it is not obvious whether the Cauchy problem is well-posed.

4

Enumerative combinatorics

Summary

In this chapter we discuss the results obtained by Fewster and the author in [95] on the enumeration of the run structures of permutations. Some of the results stated here will can be applied in the study of the moment problem in quantum field theory and the connection will be discussed briefly in Sect. 4.4.

The first section (Sect. 4.1) gives a summary of the elementary definitions for (linear) permutations and circular permutations. Then, the subsections of the second section (Sects. 4.2.1 to 4.2.3) deal, respectively, with the enumeration of the run structure of atomic, circular and linear permutations. Using a suitable decomposition, this is accomplished in each case by reducing the enumeration problem to that for atomic permutations. In the third section (Sect. 4.3) we apply and extend the methods developed in the preceeding sections to enumerate the valleys of permutations, thereby reproducing a result of Kitaev [138]. Finally, in the last section (Sect. 4.4), we discuss the original motivation of the work [95] and other possible applications.

4.1 Permutations

Let us adopt the following notation for integer intervals: $[a..b] \doteq [a, b] \cap \mathbb{N} = \{a, a + 1, \ldots, b\}$ with the special case $[n] \doteq [1..n]$.

4.1.1 Linear permutations

Given a set S, a *(linear) permutation* of S is a bijection $\sigma : S \to S$. In the *two-line notation* of the permutation of a finite set is written as

$$\sigma = \begin{pmatrix} a & b & c & \cdots \\ \sigma(a) & \sigma(b) & \sigma(c) & \cdots \end{pmatrix},$$

where $a, b, c, \ldots \in S$. It is clear that the order of elements in the first line is irrelevant as long as the second line is ordered accordingly.

The set of all bijection on S forms the *(linear) permutation group* $\mathfrak{S}_S$ of S; the group operation is the composition $\circ$ of functions. There are $n!$ permutations in $\mathfrak{S}_S$ if S is a set of n elements. In the special case that $S = [n]$, one writes $\mathfrak{S}_n \doteq \mathfrak{S}_{[n]}$.

Given a (strict) total order on a finite set S, *i.e.*, a binary relation $<$ that is transitive and trichotomous, the first line will always be ordered in the natural order. Since every finite ordered set S is isomorphic to a subset $[n]$ of the natural numbers with the standard ordering, this identification will tacitly be assumed henceforth.

Therefore the first line can be disposed of and one can use instead the *one-line notation*

$$\sigma = \begin{pmatrix} \sigma(1) & \sigma(2) & \sigma(3) & \cdots \end{pmatrix}.$$

We see that a permutation is equivalent to a change of the linear order of the set S. Further condensing the one-line notation, the permutations of a finite ordered set can be identified with words

$$\sigma = \sigma_1 \sigma_2 \sigma_3 \cdots,$$

where the shorthand $\sigma_i = \sigma(i)$, $i \in S$, was used.

4.1.2   Circular permutations

Instead of considering different orderings of a set along a line, one can study different arrangements of the elements of the set on an oriented circle (turning the circle over produces in general a different permutation).

Let S be a set with a distinguished element e. The *circular permutations* of S are the bijections $\sigma : S \to S$ that preserve e, *i.e.*, $\sigma(e) = e$. The circular permutations of S also form a group, the *circular permutation group* $\mathfrak{C}_S$; if $S = [n]$, define $\mathfrak{C}_n \doteq \mathfrak{C}_{[n]}$. Clearly, $\mathfrak{C}_S$ is a subgroup of $\mathfrak{S}_S$ and its cardinality is $(n-1)!$ if that of S is n.

A *cyclic order* is a ternary relation $[\cdot, \cdot, \cdot]$ on a set S is a set of triples $T \subset S^{\times 3}$ that satisfies

(a) $[a, b, c] \in T$ implies $[b, c, a] \in T$ (cyclicity),

(b) $[a, b, c] \in T$ implies $[c, b, a] \notin T$ (asymmetry),

(c) $[a, b, c], [a, c, d] \in T$ implies $[a, b, d] \in T$ (transitivity),

(d) a, b, c mutually distinct implies either $[a, b, c] \in T$ or $[c, b, a] \in T$ (totality).

Every (strict) total order $<$ induces a cyclic order by setting $[a, b, c] \in T$ if and only if $a < b < c$ or $b < c < a$ or $c < b < a$. Conversely, every cyclic order induces different possible linear orders. Namely, setting $a < b$ if and only if $[a, b, e]$ for fixed $e \in S$ yields a total order on $S \setminus \{e\}$ which can be extended to a linear order on S by defining e as either the minimal or maximal element of the set. Consequently, the natural choice for the distinguished element of a finite ordered set S in the construction above is the minimal or maximal element of S; we will always choose the minimal element.

The different notations for linear permutations generalize straightforwardly to circular permutations. Given a circular permutation σ of a finite ordered set S, write

$$\sigma = \begin{pmatrix} 1 & \sigma(2) & \sigma(3) & \cdots \end{pmatrix}.$$

To distinguish circular permutations more clearly from linear ones and to highlight the circular symmetry, we modify the word-notation in the case of a circular permutation σ to

$$\sigma = \dot{1}\sigma_2\sigma_3 \cdots \dot{\sigma}_n,$$

in analogy with the notation for repeating decimals when representing rational numbers. Moreover, for convenience we define $\sigma(n + 1) \doteq \sigma(1)$ for all circular permutations σ of n-element sets.

4.1.3 Atomic permutations

Let us introduce a special subgroup of $\mathfrak{S}_S$ for a finite ordered set S with minimal element e maximal element m.

Definition 4.1. *Define the* rising atomic permutations[1] $\mathfrak{A}_S^+ \subset \mathfrak{S}_S$ *as those permutations that satisfy $\sigma(e) = e$ and $\sigma(m) = m$ for all $\sigma \in \mathfrak{A}_S^+$. The* falling atomic permutations $\sigma \in \mathfrak{A}_S^-$ *are the reversed rising atomic permutations, i.e., $\sigma(e) = m$ and $\sigma(m) = e$.*

Naturally, the cardinality of $\mathfrak{A}_S^\pm$ is $(n - 2)!$. If $S = [n]$, we write $\mathfrak{A}_n^+$ ($\mathfrak{A}_n^-$) and see that it is the set of permutations of the form $1 \cdots n$ ($n \cdots 1$).

Let us discuss the significance of the atomic permutations. We say that a permutation $\sigma \in \mathfrak{S}_S$ of S *contains an atomic permutation* $\pi \in \mathfrak{A}_T$, $T \subset S$, if π can be considered a subword of σ. The atomic permutation π in σ is called *inextendible* if σ contains no other atomic permutation $\pi' \in \mathfrak{A}_{T'}$, $T' \subset S$, such that $T \subsetneq T'$.

In particular, any permutation $\sigma \in \mathfrak{S}_S$ of S with $|S| \geq 2$ contains an inextendible atomic permutation $\pi \in \mathfrak{S}_T$ of a subset $T \subset S$ that contains both the smallest and the largest element of S. That is, if $S = [n]$ and we consider σ as a word, it contains a subword π of the form $1 \cdots n$ or $n \cdots 1$. The permutation π will be called the *principal atom* of σ.

Proposition 4.2. *Any permutation $\sigma \in \mathfrak{S}_S$ of a finite set $S \subset \mathbb{N}$ can be uniquely decomposed into a tuple $(\pi^1, \ldots, \pi^k)$ of inextendible atomic permutations $\pi^i \in \mathfrak{A}_{T_i}$, $T_i \subset S$ (non-empty) such that $\pi^i_{|T_i|} = \pi^{i+1}_1$ for all $i < k$ and $\cup_i T_i = S$. We call π^i the* atoms *of σ.*

Proof. *Existence:* It is clear that any permutation of a set of 1 or 2 elements is an atomic permutation. Suppose, for some $n \geq 3$, that all permutations of $n - 1$ elements or less can be decomposed into inextendible atomic permutations. Without loss of generality, we show that any non-atomic permutation $\sigma \in \mathfrak{S}_n$ also has a decomposition into inextendible atomic permutations. Regarding σ as a word, we can write $\sigma = \alpha \cdot n \cdot \omega$, where α and ω are non-empty subwords. Notice that the permutations $\alpha \cdot n$ and $n \cdot \omega$ have a unique decomposition by assumption. Since an atomic permutation begins or ends with the largest element, we find that a decomposition of σ into inextendible atomic permutations is given by the combination of the decompositions of $\alpha \cdot n$ and $n \cdot \omega$.

Uniqueness: This is clear from the definition of inextendibility.   □

Because of this property, the atomic permutations will prove to be very useful.

---

[1]The rationale for this naming should become clear later when we see that arbitrary permutations can be decomposed into atomic permutations, but no further.

4.1.4   Mountaineering

Given a (linear or circular) permutation σ of an ordered set S of cardinality n, a position $i < n$ is a *descent* of σ if $\sigma(i) > \sigma(i+1)$. Any $i < n$ of a permutation σ that is not a descent is called an *ascent* of σ. For example, the permutation 52364178 has the descents $1, 4, 5$ and the ascents $2, 3, 6, 7$, whereas the circular permutation $\dot{1}4536782\dot{2}$ has the descents $3, 7, 8$ and the ascents $1, 2, 4, 5, 6$.

All the descents of a (linear or circular) permutation σ can be collected in the *descent set*

$$D(\sigma) \doteq \{i \mid i \text{ is a descent of } \sigma\}.$$

It is an elementary exercise in enumerative combinatorics to count the number of linear permutations of $[n]$ whose descent set is given by a fixed $S \subseteq [n-1]$. Let $S = \{s_1, s_2, \ldots, s_k\}$ be an ordered subset of $[n-1]$, then [37, Thm. 1.4]

$$\beta(S) \doteq \left|\{\sigma \in \mathfrak{S}_n \mid D(\sigma) = S\}\right| = \sum_{T \subseteq S}(-1)^{|S-T|}\binom{n}{s_1, s_2 - s_1, s_3 - s_2, \ldots, n - s_k}.$$

This result can also be adapted to circular permutations.

Related to the notions of ascents and descents are the concepts of *peaks* and *valleys*. A peak occurs at position $i \in [2..n-1]$ of a linear permutation σ if $\sigma(i-1) < \sigma(i) > \sigma(i+1)$, whereas a valley occurs in the opposite situation $\sigma(i-1) > \sigma(i) < \sigma(i+1)$. Again, this notion can be generalized to circular permutations, where, additionally, 1 is always a valley and n is a peak if and only if $\sigma(n) > \sigma(n-1)$. In the example above, 4 is a peak $2, 6$ are valleys of 52364178, whereas $3, 7$ are peaks and 1 is a valley for $\dot{1}4536782\dot{2}$, see also Fig. 4.1.

4.2   Run structures

Definition 4.3. *A* run *r of a (linear or circular) permutation σ is an interval $[i..j]$ such that $\sigma(i) \gtrless \sigma(i+1) \gtrless \cdots \gtrless \sigma(j)$ is a monotone sequence, either increasing or decreasing, and so that it cannot be extended in either direction; its length is defined to be $j - i$. If σ is a permutation of an n-element set, the collection of the lengths of all runs gives a partition p of $n - 1$ (linear permutations) or n (circular permutations). The partition p is called the* run structure *of σ.*

It follows that a run starts and ends at peaks, valleys or at the outermost elements of a permutation. For example, the permutation 52364178 has runs $[1..2]$, $[2..4]$, $[4..6]$, $[6..8]$ with lengths $1, 2, 2, 2$, whereas the circular permutation $\dot{1}4536782\dot{2}$ has runs $[1..3]$, $[3..4]$, $[4..7]$, $[7..9]$, of lengths $2, 1, 3, 2$. Representing these runs by their image under the permutation, they are more transparently written as 52, 236, 641, 178 and 145, 53, 3678, 821 respectively. The runs of permutations can also be neatly represented as directed graphs as shown in Fig. 4.1. In these graphs the peaks and valleys correspond to double sinks and double sources.

Motivated by a problem in mathematical physics [90] (see also Sect. 4.4), we are interested in the following issue, which we have not found discussed in the literature. By definition, the run structure associates each permutation $\sigma \in \mathfrak{C}_n$ with a partition p of n. For example, $\dot{1}4536782\dot{2}$ and $\dot{1}3452786\dot{6}$ both correspond to the same partition
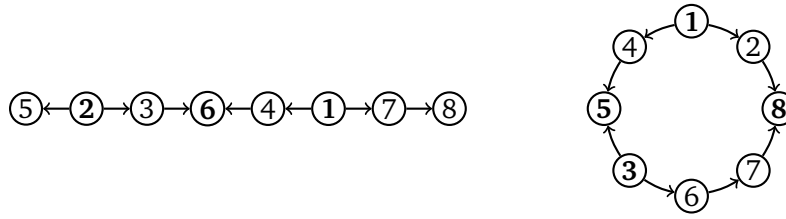
Figure 4.1. The two directed graphs representing the runs of the linear permutation 52364178 (left) and the circular permutation $\dot{1}4536782\dot{}$ (right). Peaks and valleys are indicated by boldface numbers.

$1 + 2 + 2 + 3$ of 8. Our interest is in the inverse problem: given a partition p of n, we ask for the number $Z_{\mathfrak{C}}(p)$ of circular permutations whose run structure is given by p. One may consider similar questions for other classes of permutations, with slight changes; for example, note that the run structure of a permutation $\sigma \in \mathfrak{S}_n$ is a partition of $n - 1$.

To put the research in [95] in perspective with the existing literature on the enumerative cominatorics of permutations, a short remark is in order: The enumeration of permutations according to their run structure was already discussed by André [16] for alternating permutations, *i.e.*, permutations that alternate between ascents and descents. In [42] the enumeration of linear permutations according to the order and length of their runs was studied, so obtaining a map to compositions, rather than partitions. In contrast to this approach, the method discussed in [95] was designed to facilitate computation; for the application in [90] calculations were taken up to 65 runs using exact integer arithmetic in Maple$^{\text{TM}}$ [152].

4.2.1 Atomic permutations

We now begin the enumeration of atomic permutations according to their run structure. That is, for every partition p of $n - 1$ we aim to find the number $Z_{\mathfrak{A}}(p)$ of atomic permutations $\mathfrak{A}_n^{\pm}$ of length n.

Observe that any $\sigma \in \mathfrak{A}_n^+$ can be extended to a permutation in $\mathfrak{A}_{n+1}^+$ by replacing n with $n + 1$ and reinserting n in any position after the first and before the last. Thus, 13425 can be extended to 153426, 135426, 134526 or 134256. Every permutation in $\mathfrak{A}_{n+1}^+$ arises in this way, as can be seen by reversing the procedure. The effect on the run lengths can be described as follows.

Case 1: The length of one of the runs can be increased by one by inserting n either at

1. the end of an increasing run if it does not end in $n + 1$, thereby increasing its length (*e.g.*, $13425 \rightarrow 134526$)
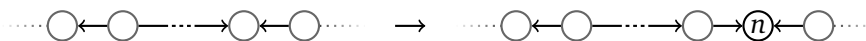
2. the penultimate position of an increasing run, thereby increasing its own length if it ends in $n + 1$ (*e.g.*, $13425 \rightarrow 135426$) or increasing the length of the following decreasing run otherwise (*e.g.*, $13425 \rightarrow 134256$)
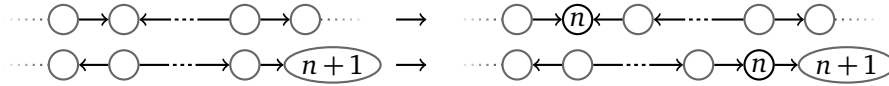
Case 2: Any run of length $i + j \geq 2$ becomes three run of lengths 1, i and j if we insert n either after

1. i elements of an increasing run (*e.g.*, $13425 \to 153426$ exemplifies $i = 1$, $j = 1$)

2. $i + 1$ elements of a decreasing run (*e.g.*, $14325 \to 143526$ for $i = 1$, $j = 1$)

An analogous argument can be made for the falling atomic permutations $\mathfrak{A}_n^-$.

Notice that every partition of a positive integer n can be represented by monomials in the ring of polynomials[2] $\mathbb{Z}[x_1, x_2, \ldots, x_n]$. Namely, we can express a partition $p = p_1 + p_2 + \cdots + p_k$ as $x_{p_1} x_{p_2} \cdots x_{p_k}$ (for example, the partition $1 + 2 + 2 + 3$ of 8 is written as $x_1 x_2^2 x_3$).

Now, let p be a partition and X the corresponding monomial. To this permutation there correspond $Z_{\mathfrak{A}}(p)$ permutations in $\mathfrak{A}_n^\pm$ which can be extended to permutations in $\mathfrak{A}_{n+1}^\pm$ in the manner described above. Introducing the (formally defined) differential operator

$$\mathscr{D} \doteq \mathscr{D}_0 + \mathscr{D}_+ \quad \text{with} \quad \mathscr{D}_0 \doteq \sum_{i=1}^{\infty} x_{i+1} \frac{\partial}{\partial x_i}, \quad \mathscr{D}_+ \doteq \sum_{i,j \geq 1} x_1 x_i x_j \frac{\partial}{\partial x_{i+j}}, \qquad (4.1)$$

we can describe this extension in terms of the action of $\mathscr{D}$ on X. We say that $\mathscr{D}_0$ is the *degree-preserving* part of $\mathscr{D}$; it represents the case 1 of increasing the length of a run: the differentiation $\partial / \partial x_i$ removes one of the runs of length i and replaces it by a run of length $i + 1$, keeping account of the number of ways in which this can be done. Similarly, case 2 of splitting a run into 3 parts is represented by the *degree-increasing* part $\mathscr{D}_+$. For example, each of the 7 atomic permutations corresponding to the partition $1 + 1 + 3$ can be extended as

$$\mathscr{D} x_1^2 x_3 = 2 x_1 x_2 x_3 + x_1^2 x_4 + x_1^4 x_2,$$

i.e., each can be extended to two atomic permutations corresponding to the partitions $1 + 2 + 3$, one corresponding to $1 + 1 + 4$ and one to $1 + 1 + 1 + 1 + 2$.

Therefore, starting from the trivial partition 1 of 1, represented as x_1, we can construct a recurrence relation for polynomials $A_n = A_n(x_1, x_2, \ldots, x_n)$ which, at every step $n \geq 1$, encode the number of atomic permutations $Z_{\mathfrak{A}}(p)$ of length $n + 1$

---

[2]If one wants to encode also the order of the run (*e.g.*, to obtain a map from permutations of length n to the compositions of n), one can exchange the polynomial ring with a noncommutative ring. Alternatively, if one wants to encode the direction of a run, one could study instead the ring $\mathbb{Z}[x_1, y_1, x_2, y_2, \ldots]$, where x_i denotes an increasing run of length i and y_j encodes a decreasing run of length j.

with run structure given by a partition p of n as the coefficients of the corresponding monomial in A_n. The polynomial A_n, accordingly defined by

$$A_n = \sum_{p \vdash n} Z_{\mathfrak{A}}(p) \prod_{i=1}^{n} x_i^{p(i)}, \tag{4.2}$$

where the sum is over all partitions p of n and $p(i)$ denotes the multiplicity of i in the partition p, can thus be computed from the recurrence relation

$$A_1 \doteq x_1, \tag{4.3a}$$

$$A_n \doteq \mathscr{D}A_{n-1}, \qquad (n \geq 2). \tag{4.3b}$$

We say that the polynomials A_n enumerate the run structure of the atomic permutations.

We summarize these results in the following proposition:

Proposition 4.4. *The number $Z_{\mathfrak{A}}(p)$ of rising or falling atomic permutations of length $n - 1$ corresponding to a given run structure (i.e., a partition p of n), is determined by the polynomial A_n via (4.2). The polynomials A_n satisfy the recurrence relation (4.3).*

Note that atomic permutations always contain an odd number of runs and thus $Z_{\mathfrak{A}}(p)$ is zero for even partitions p.

It will prove useful to combine all generating functions A_n into the formal series

$$\mathscr{A}(\lambda) \doteq \sum_{n=0}^{\infty} A_{n+1} \frac{\lambda^n}{n!} = \sum_{n=0}^{\infty} \mathscr{D}^n A_1 \frac{\lambda^n}{n!},$$

which can be expressed compactly as the exponential

$$\mathscr{A}(\lambda) = \exp(\lambda \mathscr{D}) A_1.$$

The first few A_n are given by

$$A_2 = x_2$$
$$A_3 = x_3 + x_1^3$$
$$A_4 = x_4 + 5x_2 x_1^2$$
$$A_5 = x_5 + 7x_3 x_1^2 + 11x_2^2 x_1 + 5x_1^5$$
$$A_6 = x_6 + 9x_4 x_1^2 + 11x_2^3 + 38x_3 x_2 x_1 + 61x_2 x_1^4.$$

from which we can read off that there is 1 permutation in $\mathfrak{A}_6^{\pm}$ corresponding to the trivial partition $5 = 5$, 7 corresponding to the partition $5 = 1+1+3$, 11 corresponding to $5 = 1+2+2$ and 5 corresponding to $5 = 1+1+1+1+1$. As a check, we note that $1 + 7 + 11 + 5 = 24$, which is the total number of elements of $\mathfrak{A}_6^{\pm}$; similarly, the coefficients in the expression for A_6 sum to 120, the cardinality of $\mathfrak{A}_7^{\pm}$. A direct check that the coefficients in A_n sum to $(n-1)!$ for all n will be given in the last paragraph of Sect. 4.3.

The first degree term $A_n^{(1)}$ of A_n is x_n as can be seen by a trivial induction using $A_n^{(1)} = \mathscr{D}_0 A_{n-1}^{(1)}$, which follows from the recurrence relation (4.3). Therefore $Z_{\mathfrak{A}}(n) = 1$.

For $A_n^{(k)}$ with $k > 1$ also the effect of $\mathscr{D}_+$ has to be taken into account, complicating things considerably. Nevertheless, the general procedure is clear: once $A_m^{(k-2)}$ is known for all $m < n$, $A_n^{(k)}$ can be obtained as

$$A_n^{(k)} = \mathscr{D}_0 A_{n-1}^{(k)} + \mathscr{D}_+ A_{n-1}^{(k-2)} = \sum_{m=k-1}^{n-1} \mathscr{D}_0^{n-m-1} \mathscr{D}_+ A_m^{(k-2)}.$$

Here one can make use of the following relation. Applying $\mathscr{D}_0$ repeatedly to any monomial $x_{i_1} x_{i_2} \cdots x_{i_k}$ of degree k yields, as a consequence of the Leibniz rule,

$$\mathscr{D}_0^n x_{i_1} x_{i_2} \cdots x_{i_k} = \sum_{\substack{j_1, j_2, \ldots, j_k \geq 0 \\ j_1 + j_2 + \cdots + j_k = n}} \binom{n}{j_1, j_2, \ldots, j_k} x_{i_1 + j_1} x_{i_2 + j_2} \cdots x_{i_k + j_k}. \qquad (4.4)$$

This observation provides the means to determine the third degree term $A_n^{(3)}$. Applying $\mathscr{D}_+$ to any $A_m^{(1)} = x_m$ with $m \geq 2$ produces $x_1 x_p x_q$ with $p + q = m$ and $p, q \geq 1$. Moreover, the repeated action of $\mathscr{D}_0$ on $x_1 x_p x_q$ is described by (4.4) and thus

$$A_n^{(3)} = \sum_{\substack{p,q,r,s,t \geq 0 \\ 1+p+q+r+s+t=n}} \binom{n-p-q-1}{r,s,t} x_{1+r} x_{p+s} x_{q+t}.$$

After some algebra this yields

Proposition 4.5. *The third degree term $A_n^{(3)}$ of the polynomial $A_n, n \geq 3$, is given by*

$$A_n^{(3)} = \sum_{\substack{i,j,k \geq 1 \\ i+j+k=n}} \sum_{q=1}^{k} \frac{n-q-1}{n-q-j} \binom{n-q-2}{i-1, j-1, k-q} x_i x_j x_k. \qquad (4.5)$$

The equation (4.5) for the third degree term $A_n^{(3)}$ can be rewritten into a formula for $Z_{\mathfrak{A}}(p_1 + p_2 + p_3)$, *i.e.*, the number of permutations of $[n+1]$ that start with 1, end with $n+1$ and have three runs of lengths p_1, p_2, p_3, by changing the first sum to a sum over $i, j, k \in \{p_1, p_2, p_3\}$. In particular, this gives rise to three integer series for the special cases

$$Z_{\mathfrak{A}}(n + n + n), \quad Z_{\mathfrak{A}}(1 + n + n), \quad Z_{\mathfrak{A}}(1 + 1 + n),$$

with $n \in \mathbb{N}$.

The first series

$$Z_{\mathfrak{A}}(n + n + n) = \sum_{q=1}^{n} \frac{3n-q-1}{2n-q} \binom{3n-q-2}{n-1, n-1, n-q}$$

$$= 1, 11, 181, 3499, 73501, 1623467, \ldots \qquad (n \geq 1)$$

gives the number of atomic permutations with three runs of equal length n. It does not appear to be known in the literature nor can it be found in the OEIS [164] and the existence of closed form expression is currently unkown. For the second series,

however, a simple closed form can be found:

$$Z_{\mathfrak{A}}(1+n+n) = \sum_{q=1}^{n} \left(\binom{2n-q}{n-1} + \binom{2n-q-1}{n-1} \right) + \frac{1}{2} \binom{2n}{n}$$

$$= 2 \binom{2n}{n} - 1 = 11, 39, 139, 503, 1847, \ldots, \qquad (n \geq 2)$$

is the number of atomic permutations in $\mathfrak{A}_{2n+2}^{\pm}$ with two runs of length n. One may understand this directly: there are $\binom{2n}{n}$ permutations in which the length 1 run is between the others and $\binom{2n}{n} - 1$ in which it is either first or last. The third series, $Z_{\mathfrak{A}}(1+1+n)$, i.e., the number of atomic permutations in $\mathfrak{A}_{n+3}^{\pm}$ with two runs of length 1, is given by the odd numbers bigger than 3:

$$Z_{\mathfrak{A}}(1+1+n) = 2n+1 = 5, 7, 9, 11, 13, 15, \ldots, \qquad (n \geq 2).$$

Observe that terms of the form x_1^n in A_n encode alternating permutations, which were already investigated by André in the 1880's [17]. As a consequence of his results, we find that the alternating atomic permutations are enumerated by the *secant numbers* S_n, the coefficients of the Maclaurin series of $\sec x = S_0 + S_1 x^2/2! + S_2 x^4/4! + \cdots$,

$$Z_{\mathfrak{A}}\left(\sum_{i=1}^{2n+1} 1 \right) = S_n = 1, 1, 5, 61, 1385, 50521, \ldots \quad (n \geq 0, \text{ OEIS series A000364}).$$

This is due to the fact that all alternating atomic permutations of $[2n]$ can be understood as the reverse alternating permutations of $[2..2n-1]$ with a prepended 1 and an appended $2n$. Moreover, since any x_1^{2n+1} can only be produced through an application of $\mathscr{D}$ on $x_2 x_1^{2(n-1)}$, we also have $Z_{\mathfrak{A}}\left(2 + \sum_{i=1}^{2(n-1)} 1 \right) = S_n$.

4.2.2 Circular permutations

The methods developed in the last section to enumerate atomic permutations can also be applied to find the number of circular permutations $Z_{\mathfrak{C}}(p)$ with a given run structure p. Indeed, any circular permutation in $\mathfrak{C}_{n-1}$ can be extended to a permutation in $\mathfrak{C}_n$ by inserting n at any position after the first (e.g., $\dot{1}453\dot{2}$ can be extended to $\dot{1}6453\dot{2}$, $\dot{1}4653\dot{2}$, $\dot{1}4563\dot{2}$, $\dot{1}4536\dot{2}$ or $\dot{1}4532\dot{6}$). As in the case of atomic permutations, this extension either increases the length of a run or splits a run into three runs. Namely, we can increase the length of one run by inserting n at the end or the penultimate position of an increasing run or we can split a run of length $i + j \geq 2$ into three runs of lengths i, j and 1 by inserting n after i elements of an increasing run or after $i + 1$ elements of a decreasing run.

We introduce polynomials C_n representing the run structures of all elements of $\mathfrak{C}_n$, by analogy with the polynomials A_n in the previous section:

$$C_n = \sum_{p \vdash n} Z_{\mathfrak{C}}(p) \prod_{i=1}^{n} x_i^{p(i)} \tag{4.6}$$

and we say that the polynomials C_n enumerate the run structure of the circular permutations. In the last paragraph we saw that we can use the differential operator

$\mathscr{D}$ introduced in (4.1) to find a recurrence relation similar to (4.3). Namely,

$$C_2 \doteq x_1^2, \tag{4.7a}$$

$$C_n \doteq \mathscr{D}C_{n-1}, \qquad (n \geq 3) \tag{4.7b}$$

giving in particular

$$C_3 = 2x_2 x_1$$
$$C_4 = 2x_2^2 + 2x_3 x_1 + 2x_1^4$$
$$C_5 = 2x_4 x_1 + 6x_3 x_2 + 16x_1^3 x_2$$
$$C_6 = 2x_5 x_1 + 8x_4 x_2 + 6x_3^2 + 62x_1^2 x_2^2 + 26x_1^3 x_3 + 16x_1^6$$

from which we can read off that there are 2 permutations in $\mathfrak{C}_5$ corresponding to $5 = 4 + 1$, 6 corresponding to the partition $5 = 3 + 2$ and 16 corresponding to $5 = 2 + 1 + 1 + 1$. As a check, we note that $6 + 16 + 2 = 24$, which is the total number of elements of $\mathfrak{C}_5$; similarly, the coefficients in the expression for C_6 sum to 120, the cardinality of $\mathfrak{C}_6$. More on this can be found in the last paragraph of Sect. 4.3.

In summary, we have a result analogous to Prop. 4.4:

Proposition 4.6. *The number $Z_{\mathfrak{C}}(p)$ of circular permutations of length n corresponding to a given run structure p is determined by the polynomial C_n via (4.6). The polynomials C_n satisfy the recurrence relation (4.7).*

Note that circular permutations, exactly opposite to atomic permutations, always contain an even number of runs and thus $Z_{\mathfrak{C}}(p)$ is zero for odd partitions p.

The enumeration of circular and atomic permutations is closely related. In fact, introducing a generating function $\mathscr{C}$ as the formal series

$$\mathscr{C}(\lambda) \doteq \sum_{n=0}^{\infty} C_{n+2} \frac{\lambda^n}{n!} = \sum_{n=0}^{\infty} \mathscr{D}^n C_2 \frac{\lambda^n}{n!} = \exp(\lambda \mathscr{D})C_2,$$

one can show the following:

Proposition 4.7. *The formal power series $\mathscr{C}$ is the square of a formal series $\mathscr{A}$; namely,*

$$\mathscr{C}(\lambda) = \mathscr{A}(\lambda)^2 = \big(\exp(\lambda \mathscr{D})A_1 \big)^2, \tag{4.8}$$

where $A_1 \doteq x_1$.

Proof. This may be seen in various ways, but the most convenient is to study the first-order partial differential equation (in infinitely many variables)

$$\frac{\partial \mathscr{C}}{\partial \lambda} - \mathscr{D}\mathscr{C} = 0, \quad \mathscr{C}(0) = C_2 \tag{4.9}$$

satisfied by $\mathscr{C}$.

We can now apply the method of characteristics to this problem. Since it has no inhomogeneous part, the p.d.e. (4.9) asserts that $\mathscr{C}$ is constant along its characteristics. So, given λ and $x_1, x_2, \ldots$, let $\chi_1(\mu), \chi_2(\mu), \ldots$ be solutions to the characteristic

equations with $\chi_r(\lambda) = x_r$, i.e., $\chi_1(\mu), \chi_2(\mu), \dots$ are the characteristic curves which emanate from the point $(\lambda, x_1, x_2, \dots)$. Then,

$$\mathscr{C}(\lambda)|_{x_\bullet} = \mathscr{C}(0)|_{\chi_\bullet(0)} = C_2(\chi_1(0)) = \chi_1(0)^2.$$

Applying the same reasoning again to $\mathscr{A}$, which obeys the same p.d.e. as $\mathscr{C}$ but with initial condition $\mathscr{A}(0) = A_1$,

$$\mathscr{A}(\lambda)|_{x_\bullet} = \mathscr{A}(0)|_{\chi_\bullet(0)} = A_1(\chi_1(0)) = \chi_1(0).$$

Therefore, Prop. 4.7 follows by patching these two equations together. $\qquad\square$

As a consequence also the polynomials A_n and C_n are related via

$$C_n = \sum_{m=1}^{n-1} \binom{n-2}{m-1} A_m A_{n-m}. \tag{4.10}$$

It then follows that the second degree part of C_n is given by

$$C_n^{(2)} = \sum_{m=1}^{n-1} \binom{n-2}{m-1} x_m x_{n-m}$$

and, applying (4.5), that the fourth degree part can be written as

$$C_n^{(4)} = \sum_{\substack{i,j,k,l \geq 1 \\ i+j+k+l=m}} \sum_{q=1}^{k} 2 \frac{n-l-q-1}{n-l-q-j} \binom{n-2}{n-l-1} \binom{n-l-q-2}{i-1,j-1,k-q} x_i x_j x_k x_l.$$

Similar to the atomic permutations, we find that the alternating circular permutations satisfy (cf. [16, §41])

$$Z_{\mathfrak{C}}\left(\sum_{i=1}^{2n} 1\right) = T_n = 1, 2, 16, 272, 7936, 353792, \dots \quad (n \geq 1, \text{OEIS series A000182})$$

and also $Z_{\mathfrak{C}}(2 + \sum_{i=1}^{2n-3} 1) = T_n$, where T_n are the *tangent numbers*, the coefficients of the Maclaurin series of $\tan x = T_1 x_1 + T_2 x_3/3! + T_3 x_5/5! + \cdots$. Furthermore, from (4.10) we find the relation

$$T_{n+1} = \sum_{m=0}^{n} \binom{2n}{2m} S_m S_{n-m},$$

which can be traced back to $\tan' x = \sec^2 x$.

To conclude this section, we note that the argument of Prop. 4.7 proves rather more: namely, that $\exp(\lambda\mathscr{D})$ defines a ring homomorphism from the polynomial ring $\mathbb{C}[x_1, x_2, \dots]$ to the ring of formal power series $\mathbb{C}[[x_1, x_2, \dots]]$. This observation can be used to accelerate computations: for example, the fact that $A_3 = x_3 + x_1^3$ implies that

$$\mathscr{A}''(\lambda) = \mathscr{A}(\lambda)^3 + \exp(\lambda\mathscr{D})x_3,$$

which reduces computation of $A_{n+3} = \mathscr{D}^{n+2} x_1$ to the computation of $\mathscr{D}^n x_3$. Once $\mathscr{A}$ is obtained, we may of course determine $\mathscr{C}$ by squaring.

4.2.3   Linear permutations

In the last section we studied the run structures of circular permutations $\mathfrak{C}_n$ and discovered that their run structures can be enumerated by the polynomials A_n. One might ask, what the underlying reason for this is. Circular permutations of $[n]$ have the same run structure as the linear permutations of the multiset $\{1, 1, 2, \ldots, n\}$ which begin and end with 1. These permutations can then be split into two atomic permutations at the occurrence of their maximal element. For example, the circular permutation $\dot{1}453\dot{2}$ can be split into the two atomic permutations 145 of $\{1, 4, 5\}$ and 5321 of $\{1, 2, 3, 5\}$. This also gives us the basis of a combinatorial argument for the fact that $\mathscr{C} = \mathscr{A}^2$. Similarly it is in principle possible to encode the run structures of any subset of permutations using the polynomials A_n. The goal of this section is to show how this may be accomplished for $\mathfrak{S}_S$ for any $S \subset \mathbb{N}$.

As in Sects. 4.2.1 and 4.2.2, we want to find polynomials

$$L_n = \sum_{p \vdash n} Z_{\mathfrak{S}}(p) \prod_{i=1}^{n} x_i^{p(i)}$$

that enumerate the run structure of the permutations $\mathfrak{S}_{n+1}$. This may be achieved in a two step procedure. Since every permutation has a unique decomposition into inextendible atomic permutations, we can enumerate the set of permutations according to this decomposition. The enumeration of permutations by their run structure follows because the enumeration of atomic permutations has already been achieved in Sect. 4.2.1.

The key to our procedure is to understand the factorisation of the run structure into those of atomic permutations. Considering $\sigma \in \mathfrak{S}_n$ as a word, we can write it as the concatenation $\sigma = \alpha \cdot \pi \cdot \omega$, where π is the principal atom of σ (see Sect. 4.1.3) and α, ω are (possibly empty) subwords of σ. Since the decomposition of σ into its atoms also decomposes its run structure, the complete runs of σ are determined by the runs of $\alpha \cdot 1$, π and $n \cdot \omega$ if π is rising, or of $\alpha \cdot n$, π and $1 \cdot \omega$ if π is falling.

Let S_ω be the set of letters in ω and define $\rho : S_\omega \to S_\omega$ to be the involution mapping the i'th smallest element of S_ω to the i'th largest, for all $1 \leq i \leq |S_\omega|$. Then the run structure of $n \cdot \omega$ is identical to that of $1 \cdot \rho(\omega)$, where $\rho(\omega)$ is obtained by applying ρ letterwise to ω. Furthermore, in the case $\pi = 1 \cdots n$, the combined run structures of $\alpha \cdot 1$ and $n \cdot \omega$ are precisely the run structure of $\alpha \cdot 1 \cdot \rho(\omega)$, while, if $\pi = n \cdots 1$, the combined run structures of $\alpha \cdot n$ and $1 \cdot \omega$ precisely form the run structure of $\alpha \cdot n \cdot \rho(\omega)$. We refer to $\alpha \cdot 1 \cdot \rho(\omega)$ or $\alpha \cdot n \cdot \rho(\omega)$ as the *residual permutation*.

Summarising, the run structure of σ may be partitioned into that of π and either $\alpha \cdot 1 \cdot \rho(\omega)$ or $\alpha \cdot n \cdot \rho(\omega)$; accordingly, the monomial for σ factorises into that for the principal atom π and that for the residual permutation. Therefore, the polynomial enumerating linear permutations by run structure can be given in terms of the those enumerating atomic permutations of the same or shorter length and of linear permutations of strictly shorter length.

This argument can be used to give a recursion relation for L_n, which enumerates permutations of $[n+1]$ by their run structure. Taking into account that the principal atom consists of $m + 1$ letters, where $1 \leq m \leq n$, of which $m - 1$ may be chosen freely from the set $[2 .. n]$, and that it might be rising or falling, and that the residual

permutation may be any linear permutation on a set of cardinality $n - m + 1$, we obtain the recursion relation

$$L_n = 2 \sum_{m=1}^{n} \binom{n-1}{m-1} A_m L_{n-m}, \qquad L_0 = 1.$$

Passing to the generating function,

$$\mathscr{L}(\lambda) \doteq \sum_{n=0}^{\infty} L_n \frac{\lambda^n}{n!},$$

we may deduce that

$$\frac{\partial \mathscr{L}}{\partial \lambda} = 2 \mathscr{A}(\lambda) \mathscr{L}(\lambda). \tag{4.11}$$

Our main result in this section is:

Proposition 4.8. *The run structure of all permutations in $\mathfrak{S}_{n+1}$ is enumerated by*

$$L_n = \sum_{p \vdash n} \frac{2^{|p|}}{\operatorname{ord} p} \binom{n}{p} \prod_{i=1}^{|p|} A_{p_i}, \quad L_0 = 1, \tag{4.12}$$

where the sum is over all partitions $p = p_1 + p_2 + \cdots$ of n, |p| is the number of parts of partition, $\operatorname{ord} p$ is the symmetry order of the parts of p (e.g., for $p = 1 + 1 + 2 + 3 + 3$ we have $\operatorname{ord} p = 2!2!$) and $\binom{n}{p}$ is the multinomial with respect to the parts of p. The generating function for the L_n is

$$\mathscr{L}(\lambda) \doteq \sum_{n=0}^{\infty} L_n \frac{\lambda^n}{n!} = \exp\left(2 \int_0^{\lambda} \mathscr{A}(\mu) \, d\mu \right). \tag{4.13}$$

Proof. Equation (4.13) follows immediately from (4.11), as $\mathscr{L}(0) = 1$, whereupon Faà di Bruno's formula [166, Eq. (1.4.13)] yields (4.12). $\qquad\square$

To conclude this section, we remark that the first few L_n are given by

$L_1 = 2A_1$

$L_2 = 4A_1^2 + 2A_2$

$L_3 = 8A_1^3 + 12A_1A_2 + 2A_3$

$L_4 = 16A_1^4 + 48A_1^2A_2 + 12A_2^2 + 16A_1A_3 + 2A_4$

$L_5 = 32A_1^5 + 160A_1^3A_2 + 120A_1A_2^2 + 80A_1^2A_3 + 40A_2A_3 + 20A_1A_4 + 2A_5$

$L_6 = 64A_1^6 + 480A_1^4A_2 + 320A_1^3A_3 + 720A_1^2A_2^2 + 120A_1^2A_4 + 480A_1A_2A_3 + 120A_2^3$
$\qquad + 24A_1A_5 + 60A_2A_4 + 40A_3^2 + 2A_6.$

Expanding the A_k and writing the L_n instead in terms of x_i, we obtain from these

$$L_1 = 2x_1$$
$$L_2 = 4x_1^2 + 2x_2$$
$$L_3 = 10x_1^3 + 12x_1x_2 + 2x_3$$
$$L_4 = 32x_1^4 + 58x_1^2x_2 + 12x_2^2 + 16x_1x_3 + 2x_4$$
$$L_5 = 122x_1^5 + 300x_1^3x_2 + 142x_1x_2^2 + 94x_1^2x_3 + 40x_2x_3 + 20x_1x_4 + 2x_5$$
$$L_6 = 544x_1^6 + 1682x_1^4x_2 + 568x_1^3x_3 + 1284x_1^2x_2^2 + 138x_1^2x_4 + 556x_1x_2x_3 + 142x_2^3$$
$$\quad + 24x_1x_5 + 60x_2x_4 + 40x_3^2 + 2x_6,$$

which show no obvious structure, thereby making Prop. 4.8 that much more remarkable.

4.3   Enumeration of valleys

Instead of enumerating permutations by their run structure, we can count the number of valleys of a given (circular) permutation. Taken together, the terms C_n involving a product of $2k$ of the x_i relate precisely to the circular permutations $\mathfrak{C}_n$ with k valleys. Since any circular permutation in $\mathfrak{C}_n$ can be understood as a permutation of $[3\mathinner{.\,.}n+1]$ with a prepended 1 and an appended 2 (*cf.* beginning of Sect. 4.2.3), C_n may also be used to enumerate the valleys of ordinary permutations of $[n-1]$. Namely, terms of C_{n+1} with a product of $2(k+1)$ variables x_i relate to the permutations of $\mathfrak{S}_n$ with k valleys (*i.e.*, terms of L_{n+1} which are a product of $2k$ of the x_i).

Let $V(n,k)$ count the number of permutations of n elements with k valleys. Then we see that the generating function for $V(n,k)$ for each fixed $n \geq 1$ is

$$K_n(\kappa) \doteq \sum_{k=1}^{n} \kappa^k V(n,k) = \frac{1}{\kappa} C_{n+1}(\sqrt{\kappa},\ldots,\sqrt{\kappa})$$

and we define $K_0(\kappa) \doteq 1$. The first few K_n are

$$K_1(\kappa) = 1$$
$$K_2(\kappa) = 2$$
$$K_3(\kappa) = 4 + 2\kappa$$
$$K_4(\kappa) = 8 + 16\kappa$$
$$K_5(\kappa) = 16 + 88\kappa + 16\kappa^2$$
$$K_6(\kappa) = 32 + 416\kappa + 272\kappa^2,$$

which coincide with the results in [183]. In particular, the constants are clearly the powers of 2, the coefficients of κ give the sequence A000431 of the OEIS [164] and the coefficients of κ^2 are given by the sequence A000487. Likewise, the coefficients of κ^3 may be checked against the sequence A000517. In fact, the same polynomials appear in André's work, in which he obtained a generating function closely related to (4.14) below; see [16, §158] (his final formula contains a number of sign errors, and is given in a form in which all quantities are real for κ near 0; there is also an offset, because his polynomial $A_n(\kappa)$ is our $K_{n-1}(\kappa)$).

Proposition 4.9. *The bivariate generating function, i.e., the generating function for arbitrary n, is*

$$\mathcal{K}(v,\kappa) = \sum_{n=0}^{\infty} K_n(\kappa)\frac{v^n}{n!} = 1 + \frac{1}{\kappa}\int_0^v \mathcal{C}(\mu)|_{x_1=x_2=\cdots=\sqrt{\kappa}}\,\mathrm{d}\mu$$

and is given in closed form by

$$\mathcal{K}(v,\kappa) = 1 - \frac{1}{\kappa} + \frac{\sqrt{\kappa-1}}{\kappa}\tan\left(v\sqrt{\kappa-1} + \arctan(1/\sqrt{\kappa-1})\right). \qquad (4.14)$$

This result was found by Kitaev [138] and in the remainder of this section we will show how it may be derived from the recurrence relation (4.7) of C_n.

To this end, we first note that C_{n+1} satisfies the useful scaling relation

$$\lambda^{n+1}C_{n+1}(x_1,x_2,\ldots,x_n) = C_{n+1}(\lambda x_1, \lambda^2 x_2, \ldots, \lambda^n x_n).$$

Setting $x_i = x/\lambda = \sqrt{\kappa}$ for all i, this implies

$$\lambda^{n+1}C_{n+1}(\sqrt{\kappa},\ldots,\sqrt{\kappa}) = C_{n+1}(x,\lambda x,\ldots,\lambda^{n-1}x)$$

and we find, by inserting the recurrence relations (4.7) and applying the chain rule, that with this choice of variables

$$\frac{1}{x^2}C_{n+1}(x,\lambda x,\ldots,\lambda^{n-1}x) = \lambda x\frac{\partial C_n}{\partial x} + x^2\frac{\partial C_n}{\partial \lambda} + 2\lambda C_n.$$

Hence, in turn, $K_n(\kappa) = \kappa^{-1}C_{n+1}(\sqrt{\kappa},\ldots,\sqrt{\kappa})$ satisfies the recurrence relation

$$K_n(\kappa) = 2\kappa(1-\kappa)K'_{n-1}(\kappa) + \left(2 + (n-2)\kappa\right)K_{n-1}(\kappa) \qquad (4.15)$$

for $n \geq 2$. For the bivariate generating function $\mathcal{K}$ this, together with $K_0 = K_1 = 1$, implies the p.d.e.

$$(1-v\kappa)\frac{\partial \mathcal{K}}{\partial v} + 2\kappa(\kappa-1)\frac{\partial \mathcal{K}}{\partial \kappa} + (\kappa-2)\mathcal{K} = \kappa - 1,$$

which is to be solved subject to the initial condition $\mathcal{K}(0,\kappa) = 1$.

The above equation may be solved as follows: first, we note that there is a particular integral $1 - 1/\kappa$, so it remains to solve the homogeneous equation. In turn, using an integrating factor, the latter may be rewritten as

$$(1-v\kappa)\frac{\partial}{\partial v}\frac{\kappa\mathcal{K}}{\sqrt{\kappa-1}} + 2\kappa(\kappa-1)\frac{\partial}{\partial \kappa}\frac{\kappa\mathcal{K}}{\sqrt{\kappa-1}} = 0, \qquad (4.16)$$

for which the characteristics obey

$$\frac{\mathrm{d}v}{\mathrm{d}\kappa} = \frac{1-v\kappa}{2\kappa(\kappa-1)}.$$

Solving this equation, we find that

$$v\sqrt{\kappa-1} + \arctan\frac{1}{\sqrt{\kappa-1}} = \mathrm{const}$$

along characteristics; as (4.16) asserts that $\kappa \mathcal{K}/\sqrt{\kappa-1}$ is constant on characteristics, this gives

$$\mathcal{K}(v,\kappa) = 1 - \frac{1}{\kappa} + \frac{\sqrt{\kappa-1}}{\kappa} f\left(v\sqrt{\kappa-1} + \arctan(1/\sqrt{\kappa-1})\right)$$

for some function f. Imposing the condition $\mathcal{K}(0,\kappa) = 1$, it is plain that $f = \tan$, and we recover Kitaev's generating function (4.14).

To close this section, we note that (4.15) has the consequence that $K_n(1) = nK_{n-1}(1)$ for all $n \geq 2$ and hence that $C_{n+1}(1,\ldots,1) = K_n(1) = n!$ for such n, and indeed all $n \geq 1$, because $C_2(1,1) = K_1(1) = 1$. The generating function obeys

$$\mathcal{C}(\lambda)|_{x_\bullet=1} = \sum_{n=0}^{\infty}(n+1)!\frac{\lambda^n}{n!} = (1-\lambda)^{-2}$$

for all non-negative $\lambda < 1$ from which it also follows that

$$\mathcal{A}(\lambda)|_{x_\bullet=1} = (1-\lambda)^{-1} \tag{4.17}$$

(as $A_1(1) = 1$, we must take the *positive* square root) and hence $A_n|_{x_\bullet=1} = (n-1)!$ for all $n \geq 1$. This gives a consistency check on our results: the coefficients in the expression for A_n sum to $(n-1)!$, the cardinality of $\mathfrak{A}_{n+1}^{\pm}$, while those in C_n sum to the cardinality of $\mathfrak{C}_n$. Furthermore, inserting (4.17) into the generating function $\mathcal{L}(\lambda)$ in (4.13), we find

$$\mathcal{L}(\lambda)|_{x_\bullet=1} = \sum_{n=0}^{\infty} L_n(1,\ldots,1)\frac{\lambda^n}{n!} = (1-\lambda)^{-2},$$

and thus $L_{n+1}(1,\ldots,1) = n!$, which is the cardinality of $\mathfrak{S}_n$.

4.4   Applications

The original motivation for this work arose in quantum field theory, in computations related to the probability distribution of measurement outcomes for quantities such as averaged energy densities [90]. One actually computes the cumulants κ_n ($n \in \mathbb{N}$) of the distribution: $\kappa_1 = 0$, while for each $n \geq 2$, κ_n is given as a sum indexed by circular permutations σ of $[n]$ such that $\sigma(1) = 1$ and $\sigma(2) < \sigma(n)$, in which each permutation contributes a term that is a multiplicative function of its run structure:

$$\kappa_n = \sum_{\sigma}\Phi(\sigma)$$

where $\Phi(\sigma)$ is a product over the runs of σ, with each run of length r contributing a factor y_r. Owing to the restriction $\sigma(2) < \sigma(n)$, precisely half of the circular permutations are admitted, and so $\kappa_n = \frac{1}{2}C_n(y_1, y_2, \ldots, y_n)$. Thus the cumulant generating function is

$$\begin{aligned} W(\lambda) &\doteq \sum_{n=2}^{\infty}\kappa_n\frac{\lambda^n}{n!} = \frac{1}{2}\int_0^{\lambda}\mathrm{d}\mu\,(\lambda-\mu)\mathcal{C}(\mu)|_{x_\bullet=y_\bullet} \\ &= \frac{1}{2}\int_0^{\lambda}\mathrm{d}\mu\,(\lambda-\mu)\mathcal{A}(\mu)|_{x_\bullet=y_\bullet}^2. \end{aligned}$$

and the moment generating function is $\exp W(\lambda)$ in the usual way. This expression makes sense a formal power series, but also as a convergent series within an appropriate radius of convergence. The values of y_n depend on the physical quantity involved and the way it is averaged. In one case of interest

$$y_n = 8^n \int_{(\mathbb{R}^+)^{\times n}} \mathrm{d}k_1 \, \mathrm{d}k_2 \cdots \mathrm{d}k_n \, k_1 k_2 \cdots k_n \exp\left[-k_1 - \left(\sum_{i=1}^{n-1} |k_{i+1} - k_i| \right) - k_n \right]$$

$$= 2^n \sum_{r_{n-1}=0}^{2} \sum_{r_{n-2}=0}^{2+r_{n-1}} \cdots \sum_{r_1=0}^{2+r_2} \prod_{k=1}^{n-1} (1 + r_k)$$

$$= 2, 24, 568, 20256, 966592, \ldots \qquad (n \geq 1)$$

(the sums of products must be interpreted as an overall factor of unity in the case $n = 1$). Numerical investigation leads to a remarkable identity

$$\mathcal{A}(\lambda)|_{x_\bullet = y_\bullet} = \frac{2}{1 - 12\lambda} \qquad \text{(conjectured)}$$

with exact agreement for all terms so far computed (checked up to $n = 65$). We do not have a proof for this statement, but the conjecture seems fairly secure. For example, we have shown above that $A_5 = x_5 + 7x_3 x_1^2 + 11 x_2^2 x_1 + 5x_1^5$; substituting for x_n the values of y_n obtained above, we find $A_5 = 995328$ which coincides with the fourth order coefficient in the expansion

$$\frac{2}{1 - 12\lambda} = 2 + 24\lambda + 576\frac{\lambda^2}{2!} + 20736\frac{\lambda^3}{3!} + 995328\frac{\lambda^4}{4!} + O(\lambda^5).$$

In [90], this conjecture was used to deduce

$$\exp\left(W(\lambda) \right) = \mathrm{e}^{-\lambda/6}(1 - 12\lambda)^{-1/72} \qquad \text{(conjectured)},$$

which is the moment generating function of a shifted Gamma distribution. The other generating functions of interest, with these values for the x_k are

$$\mathcal{C}(\lambda)|_{x_\bullet = y_\bullet} = \frac{4}{(1 - 12\lambda)^2}, \qquad \mathcal{L}(\lambda)|_{x_\bullet = y_\bullet} = (1 - 12\lambda)^{-1/3} \qquad \text{(conjectured)}.$$

For example, we have $C_5 = 2x_4 x_1 + 6x_3 x_2 + 16 x_1^3 x_2 = 165888$ and $L_5 = 122 x_1^5 + 300 x_1^3 x_2 + 142 x_1 x_2^2 + 94 x_1^2 x_3 + 40 x_2 x_3 + 20 x_1 x_4 + 2x_5 = 3727360$, to be compared with the terms of order λ^3 and λ^5, respectively, in the expansions

$$\frac{4}{(1 - 12\lambda)^2} = 4 + 96\lambda + 3456\frac{\lambda^2}{2!} + 165888\frac{\lambda^3}{3!} + 995328\frac{\lambda^4}{4!} + O(\lambda^5),$$

$$(1 - 12\lambda)^{-1/3} = 1 + 4\lambda + 64\frac{\lambda^2}{2!} + 1792\frac{\lambda^3}{3!} + 71680\frac{\lambda^4}{4!} + 3727360\frac{\lambda^5}{5!} + O(\lambda^6).$$

A natural question is whether there are other sequences that can be substituted for the x_k to produce generating functions with simple closed forms. To close, we give three further examples, with the corresponding generating functions computed. The first has already been encountered in Sect. 4.3 and corresponds to the case $x_k = 1$ for all $k \in \mathbb{N}$.

The second utilizes the alternating Catalan numbers: setting

$$x_{2k+1} = \frac{(-1)^k}{k+1}\binom{2k}{k}, \quad (k \geq 0), \qquad x_{2k} = 0, \quad (k \geq 1)$$

and thus $A_{2k} = 0$, we obtain, again experimentally,

$$\mathscr{C}(\lambda) = \mathscr{A}(\lambda) = 1, \quad \mathscr{L}(\lambda) = e^{2\lambda} \qquad \text{(conjectured)}$$

with exact agreement checked up to permutations of length $n = 65$. For example, one sees easily that with $x_1 = 1$, $x_3 = -1$, $x_5 = 2$ and $x_2 = x_4 = 0$, the expressions A_k and C_k given in Sects. 4.2.1 and 4.2.2 vanish for $2 \leq k \leq 6$, and have $A_1 = C_1 = 1$, likewise, $L_k = 2^k$ for $1 \leq k \leq 6$.

Third, André's classical result on alternating permutations (*cf.* last and penultimate paragraph of Sects. 4.2.1 and 4.2.2 respectively) gives the following: setting

$$x_1 = 1 \quad \text{and} \quad x_k = 0, \quad (k \geq 2)$$

we have, using (4.8) and (4.13),

$$\mathscr{A}(\lambda) = \sec \lambda, \quad \mathscr{C}(\lambda) = \sec^2 \lambda, \quad \mathscr{L}(\lambda) = (\sec \lambda + \tan \lambda)^2.$$

It seems highly likely to us that many other examples can be extracted from the structures we have described.

Moreover, we remark that it is possible to implement a merge-type sorting algorithm, called *natural merge sort* [139, Chap. 5.2.4], based upon splitting permutations of an ordered set S into its runs, which are ordered (alternatingly in ascending and descending order) sequences $S_i \subset S$. Repeatedly merging these subsequences, one ultimately obtains an ordered sequence. For example, first, we split the permutation 542368719 into 542, 368, 71 and 9. Then, we reverse every second sequence (depending on whether the first or the second sequence is in ascending order): $542 \mapsto 245$ and $71 \mapsto 17$. Depending on the implementation of the merging in the following step, this 'reversal' step can be avoided. Last, we merge similarly to the standard merge sort: $245 \vee 368 \mapsto 234568$, $17 \vee 9 \mapsto 179$ and finally $234568 \vee 179 \mapsto 123456789$. Natural merge sort is a fast sorting algorithm for data with preexisting order. Using the methods developed above to enumerate permutations by their run structure, it is in principle possible to give average (instead of best- and worst-case) complexity estimates for such an algorithm.

II

Quantum field theory

Is the purpose of theoretical physics to be no more than a cataloging of all the things that can happen when particles interact with each other and separate? Or is it to be an understanding at a deeper level in which there are things that are not directly observable (as the underlying quantized fields are) but in terms of which we shall have a more fundamental understanding?
— Julian S. Schwinger, "Quantum Mechanics" (2001), p. 24 f.

First, in order to achieve the greatest possible generality we continue our total boycott of the canonical formalism [...].
— Bryce S. DeWitt, J. Math. Phys. 3 (1962), p. 1073.

Locally covariant quantum field theory

Summary

In this chapter we will discuss the framework of locally covariant quantum field theory. In its present form it was introduced in [46] but many of its central ideas can already be found in earlier publications. It may be understood as a generalization of the Haag–Kastler axioms [109, 110] to curved spacetimes but it also differs in some subtle points because the Haag–Kastler axioms are 'more global' (see, for example, [30, 32] on the problem of gauge theories in locally covariant QFT). A generalization of the Haag–Kastler was performed by Dimock whose work [72–74] can be understood as the foundation of modern algebraic QFT on curved spacetime. Building on the work of Dimock, the paradigm of locally covariant quantum field theory should be seen a culmination of work done by Brunetti, Fredenhagen, Hollands, Kay, Verch, Wald and others on QFT on curved spacetimes, in particular renormalization, [44, 45, 123, 124, 135, 215] around the turn of the millennium after the discovery of the microlocal spectrum condition [45, 181, 182].

After discussing some general considerations leading to algebraic and locally covariant quantum field theory in the first section (Sect. 5.1), we will introduce the the general framework of locally covariant quantum field theory in the second section (Sect. 5.2). More details on the locally covariant framework may be found in the original publication [46] or also [89, 190] among many others. This is followed by an abstract study of the Borchers–Uhlmann algebra, the commutation relations and the field equation, which will lead to the field algebra, and the Weyl algebra in the third section (Sect. 5.3), we will discuss two free bosonic fields in the locally covariant framework: the scalar field and the Proca field.

5.1 General considerations

Quantum field theory is a very complex subject which cannot easily be defined. This is partly due to the fact that quantum field theory is not so much a physical theory but rather a language to formulate theories and models. However, a more important reason is that quantum field theory, even many decades after its inception, has no clear interpretation, *e.g.*, it is often not clear what the physical objects are. Nevertheless, it can be considered one of the most successful scientific discoveries ever conceived and some predictions made by quantum field theory have been tested with astonishing precision. For example, the anomalous magnetic moment of the electron has been measured in agreement with theoretical predictions in the parts in a trillion range, see [18, 117].

The relation of QFT to other theories is schematically depicted in the diagram Fig. 5.1. In particular, QFT may be considered as a Lorentz invariant quantum

$$\begin{array}{ccc}
\boxed{\text{Classical Mechanics}} & \xrightarrow{\ \ \text{Quantization}\ \ } & \boxed{\text{Quantum Mechanics}} \\
\Big\downarrow N \to \infty & & \Big\downarrow N \to \infty \\
\boxed{\text{Classical Field Theory}} & \xrightarrow[\text{Quantization}]{\ \ } & \boxed{\text{Quantum Field Theory}}
\end{array}$$
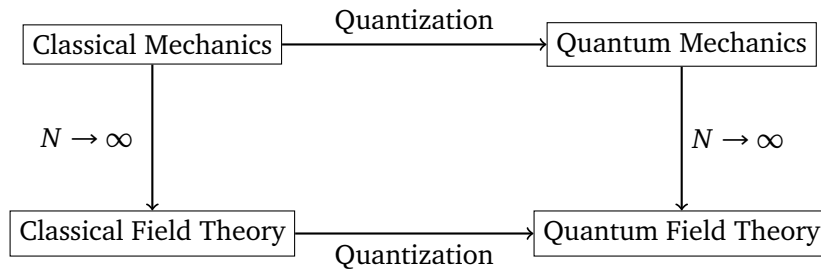
Figure 5.1.   The heuristic relation of quantum field theory with classical mechanics, classical field theory and quantum mechanics.

mechanics in the case of infinitely many degrees of freedom. One can also argue that a consistent reconciliation of quantum mechanics with special relativity (in particular locality) leads invariably to QFT, *i.e.*, fields and an infinite number of degrees of freedom are necessary, see [149] and also [52].

5.1.1   Lagrangian QFT

Quantum field theory is usually formulated in the relatively heuristic approach of the Lagrangian formalism, where, starting from a classical Lagrangian, one imposes the canonical commutation relations between the quantized position and momentum variables. In analogy to the quantum mechanical harmonic oscillator these yield creation and annihilation 'operators' on an abstract representing Hilbert space. Combining the creation and annihilation operator, one furthermore defines the quantum field. A specific Hilbert space is then selected by requiring that the annihilation operator annihilates a particular vector in the Hilbert space, the vacuum, so that one obtains the Fock space representation.

Apart from not being mathematically rigorous, the Lagrangian formalism has several conceptual drawbacks. First, it neglects *a priori* the inequivalent irreducible representations of the canonical commutation relations (as a consequence of the failure of the Stone–von Neumann theorem in infinite dimensions) and instead selects the convenient Fock space representation. However, it is not obvious what is physical and whether the inequivalent representations are simply mathematical artefacts or physically relevant. Indeed, the existence of superselection sectors shows that the presence of inequivalent representations is certainly *not* irrelevant. A closely related issue is described by Haag's theorem [110, Chap. II.1.1] which implies that the standard Fock space representation of the free theory is inequivalent to the that of the interacting theory.

Second, the fundamental entities in the Lagrangian formalism are 'operators' at a point and thus neither mathematically not physically meaningful. Physically, because it would require an infinite amount of energy to localize a field a point. Mathematically, because a field at a point is not an operator on a Hilbert space but only an operator-valued distribution. Instead one should consider field which have no sharp localization but are smeared out over a region of spacetime. That is, the fundamental entities are quantum fields smeared with compactly supported test functions.

Third, the Lagrangian formalism contains global operators, like the particle number operator, which are not operationally meaningful because they cannot be reproduced by measurements in a bounded region of spacetime. In fact, due to the Reeh–Schlieder theorem, also local number operators cannot exist [110, Thm. 5.3.2]. In this light the common interpretation of QFT in terms of localized particles is very problematic. On curved spacetimes or for accelerated observers the situation is even more problematic: as shown by the Unruh effect, the particle interpretation appears to be context dependent.

5.1.2   Algebraic QFT and locality

In the algebraic approach to quantum field theory, developed by Haag and collaborators, the problems indicated above are addressed in a conceptually simple way: Rather than taking as observables operators on a Hilbert space, in the algebraic approach one discards the concrete representation of the operators and considers only the algebraic relations satisfied between the operators. Indeed, the relations between the observables already contain a large part of the physical content of the theory.

The central pillar of algebraic quantum field theory is *locality*, better described by the German word "*Nahwirkungsprinzip*". Locality means that causally unrelated events do not influence each other and it is implemented in the following way: To every spacetime region U we can associate a local algebra of observables $\mathcal{A}(U)$ which can be measured within U. Consequently, we demand that map $U \mapsto \mathcal{A}(U)$ forms an inductive system, *i.e.*, it satisfies the *isotony* condition

$$U \subset V \implies \mathcal{A}(U) \subset \mathcal{A}(V)$$

or at least that there is an injective homomorphism $\mathcal{A}(U) \to \mathcal{A}(V)$; the correspondence $U \mapsto \mathcal{A}(U)$ for all U is called the *net of local algebras*. Further, we require that the local algebras of causally separated, causally convex regions (anti-)commute

$$U \times V \implies [\mathcal{A}(U), \mathcal{A}(V)] = \{0\}.$$

In the next section it will become clear how these conditions can be consistently imposed on different spacetimes.

Following the choice of words of Haag [110], in any concrete case, the *smeared* quantum fields may be seen as a way to 'coordinatize' the local algebras. That is, they provide a map from test functions supported a spacetime region, to the local algebra supported in that region. From this point of view it seems clear that different quantum fields can lead to equivalent algebras. The notion of quantum fields might become clearer within the categorical framework to be introduced the next section.

5.2   Framework

5.2.1   Background structure

The physical Universe appears to be well-modelled by a connected, oriented and time-oriented, four-dimensional Lorentzian manifold $(M, g, \pm, u)$, *i.e.*, a spacetime as
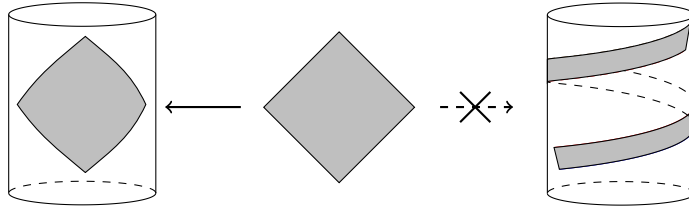
Figure 5.2. Example of an hyperbolic embedding and a non-hyperbolic embedding; see also Fig. 2.3.

defined in Sect. 2.1. Moreover, for every possible observer to carry out experiments in a finite region of spacetime one has to require that $J^+(x) \cap J^-(y)$ is compact for all $x, y \in M$. If we further require that no closed causal curves exist so that time travel is impossible, we have collected all necessary and sufficient conditions for a globally hyperbolic spacetime. Accordingly we consider as physical spacetimes all globally hyperbolic spacetimes. Of course we restrict ourselves to globally hyperbolic spacetimes also for technical reasons. In particular, only on globally hyperbolic spacetimes do we have a good understanding of the Cauchy problem for the wave equation.

To implement simultaneously *covariance* and the principle of *locality*, *i.e.*, an observer can conduct experiments in a globally hyperbolic subregion of the Universe and may remain ignorant about processes in the complement of that region, we consider as morphisms between globally hyperbolic spacetimes M, N the isometric embeddings $\psi : M \to N$ that are orientation and time-orientation preserving and whose image $\psi(M)$ is a causally convex region of N. We call these morphisms *hyperbolic embeddings*; an example and a counter-example are shown in Fig. 5.2. If the image of the hyperbolic embedding ψ contains a neighbourhood of a Cauchy surface in N, we say that it is *Cauchy*.

The set of globally hyperbolic spacetimes as objects with the hyperbolic embeddings as morphisms forms a category denoted Loc. This category was introduced in [46] and it is arguably the most fundamental but, as already mentioned in [46], not the only possible choice to describe local theories. In fact, it has been altered in the literature in various ways

- to accommodate more background structure by adding to the triple of manifold, metric and time-orientation, which is each object, additional elements like spin-structure [193], affine, principle and vector bundles [30–32] or external currents [2];

- to account for additional symmetry by allowing for more general morphisms like conformal isometries [1, 177, 5];

- to allow for the formulation of theories that are sensitive to the topology of the manifold, *e.g.*, by restricting the set of objects to manifolds that have certain de Rahm cohomology groups [1, 5].

More recently, it was suggested by Fewster [85] to consider as objects triples which instead of a metric have a global (co)frame; the morphisms are changed accordingly.
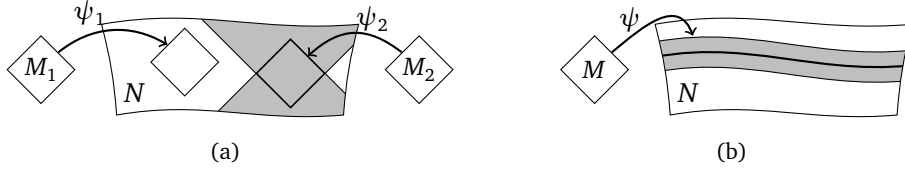
Figure 5.3. An illustration of (a) two causally disjoint embeddings and (b) a Cauchy embedding.

The resulting category is larger and encompasses the original setting via a forgetful functor but the additional structure allows for an interesting discussion of the spin-statistics theorem.

5.2.2 Observables

A theory in this categorical framework is a covariant functor from Loc into a category whose objects describe physical systems and whose morphisms encode embeddings of physical systems. In quantum field theory (on curved spacetimes) in the algebraic formulation, physical systems are modelled by $*$-algebras or C^*-algebras. Denote by *Alg the category whose objects are unital topological $*$-algebras with morphisms given by the unit-preserving $*$-monomorphisms and call a covariant functor $\mathscr{A}$: Loc $\to$ *Alg realizing such an algebra on each background in Loc a *locally covariant theory*. The algebra $\mathscr{A}(M)$ thus associated to each spacetime $M \in$ Loc is often called the *algebra of observables* although in many cases it may contain elements that are not actually physically accessible.

Given two hyperbolic embeddings $\psi_i : M_i \to N$, $i = 1, 2$, such that the images $\psi_i(M_i)$ are causally disjoint in N (*cf.* Fig. 5.3(a)), we say that $\mathscr{A}$ is *causal* if

$$[\mathscr{A}(\psi_1)\mathscr{A}(M_1), \mathscr{A}(\psi_2)\mathscr{A}(M_2)] = \{0\}.$$

Causality of $\mathscr{A}$ is closely related to its tensorial structure as discussed in [48].

The theory $\mathscr{A}$ obeys the *timeslice axiom* if

$$\mathscr{A}(\psi)\mathscr{A}(M) = \mathscr{A}(N)$$

for all Cauchy embeddings $\psi : M \to N$ (*cf.* Fig. 5.3(b)). The timeslice axiom is a prerequisite for the *relative Cauchy evolution*, which describes the response of the physical system to a perturbation of the background structure.

More concretely, let $(M, g, \pm, u)$ and $(M[h], g + h, \pm, u_h)$ be globally hyperbolic spacetimes such that h is a compactly supported symmetric tensor field and u_h is the unique time-orientation that agrees with u outside the support of h.[1] Consequently there exist neighbourhoods around two Cauchy surfaces in $M[h]$, one in the past of h and the other in the future. We can then find Cauchy morphisms $\iota^\pm$ and $\iota[h]^\pm$ from spacetimes $M^\pm \in$ Obj(Loc) into M and $M[h]$ as shown in Fig. 5.4. Together

---

[1]Note that a possible 'perturbation' is always $M[0]$ but there exists also a neighbourhood U of 0 in the set of compactly supported symmetric covariant two-tensor fields (with the test function topology) such that $(M[h], g + h)$ is globally hyperbolic for all $h \in U$, see [27, Thm. 7.2].
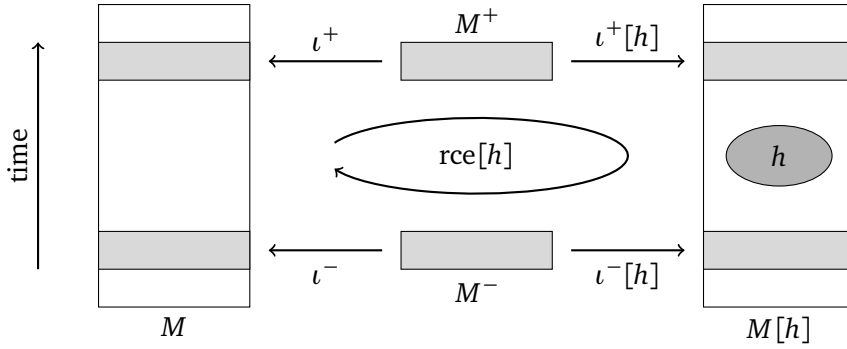
Figure 5.4. Illustration of the morphisms in the relative Cauchy evolution with the unperturbed background M on the left and the perturbed background $M[h]$ on the right.

these Cauchy morphisms make up the (*-algebra) homomorphism that is the relative Cauchy evolution map

$$\text{rce}[h] \doteq \mathscr{A}(\iota^-) \circ \mathscr{A}(\iota[h]^-)^{-1} \circ \mathscr{A}(\iota[h]^+) \circ \mathscr{A}(\iota^+)^{-1}.$$

It was shown in [46] that (in an appropriate topology, see [2, 98] for details),

$$-2\mathrm{i} \frac{\mathrm{d}}{\mathrm{d}\varepsilon} \text{rce}[\varepsilon h] A \bigg|_{\varepsilon=0} = [T(h), A]$$

for any $A \in \mathscr{A}(M)$ and where $T(h) \in \mathscr{A}(M)$ is symmetric and conserved. Since T is both symmetric and conserved it may be interpreted as a stress-energy tensor [46, 97, 98] and, in fact, in concrete models this interpretation is valid [28, 46, 93, 193].

5.3   Generalized Klein–Gordon fields

The Klein–Gordon field is usually the first field to be discussed when studying quantum field theory. We will be no different although we will perform some straightforward generalizations. Namely, we will quantize fields on natural vector bundles that satisfy a normally hyperbolic equation of motion. The results below are a generalization of those obtained in [46] for the scalar Klein–Gordon field. In principle, further generalizations of the definitions and statements presented below are possible. For example, one can replace compactly supported p-forms by compactly supported sections of 'natural' vector bundles, *i.e.*, vector bundles that, like the (co)tangent bundle, are functorially constructed from the geometric structure of the manifold. However, all these generalizations yield little insight and obfuscate some constructions. Moreover, the requirements imposed by the usual locally covariant framework make it difficult to find examples that do not just use a standard tensor bundle tensorized with a vector bundle that is independent of the geometry of the spacetime.

5.3.1   Borcher–Uhlmann algebra

For every globally hyperbolic spacetime M, let $\mathscr{D}$ be a covariant functor from Loc into the category of closed nuclear locally convex $\mathbb{C}$-vector spaces such that $\mathscr{D}(M) \subset$

$\Omega_0^p(M,\mathbb{C})$ (for fixed p and with the subspace topology) and $\mathscr{D}(\psi) = \psi_*$.

On each spacetime M we can define a straightforward generalization of the *Borchers–Uhlmann algebra* [38, 39, 213] as the unital topological *-algebra

$$\mathscr{U}(M) \doteq \bigoplus_{n\in\mathbb{N}_0} \mathscr{D}(M)^{\widehat{\otimes}n}$$

with, *i.e.*, the set of tuples $(f_n)_{n\in\mathbb{N}_0}$ with $f_n \in \mathscr{D}(M)^{\widehat{\otimes}n}$ such that only a finite number of f_n is nonzero, together with

(a) addition and scalar multiplication is component-wise,

(b) multiplication given by the canonical isomorphic embedding

$$\mathscr{D}(M)^{\widehat{\otimes}m} \otimes \mathscr{D}(M)^{\widehat{\otimes}n} \longrightarrow \mathscr{D}(M)^{\widehat{\otimes}(m+n)}$$

and extends (anti-)linearly to all of $\mathscr{U}(M)$ via the canonical embeddings of these spaces into $\mathscr{U}(M)$,

(c) a *-operation that acts on $(f_n) \in \mathscr{U}(M)$ as $(f_n)^* = (f_n^*)$ and

$$f_n^*(x_1,\dots,x_n) = \overline{f_n}(x_n,\dots,x_1),$$

(d) a topology given by the direct sum topology of the test function topology on each $\mathscr{D}(M)^{\widehat{\otimes}n}$.[2]

Assigning to each globally hyperbolic spacetime the Borchers–Uhlmann algebra $\mathscr{U}(M)$, we obtain a covariant functor $\mathscr{U} : \mathsf{Loc} \to {}^*\mathsf{Alg}$ that maps each object to the algebra and each morphism to the *-algebra morphism generated by the natural pushforward ψ_*, *i.e.*,

$$\mathscr{U}(\psi)(f_n) = (\psi_* f_n) \quad \text{with} \quad \psi_* f_n = f_n \circ (\psi^{-1})^{\otimes n}$$

for all $(f_n) \in \mathscr{U}(M)$.

Considering $\mathscr{D}(M)$ as a topological *-algebra where the involution is complex conjugation, we can consider it as a functor from Loc to $^*\mathsf{Alg}$. The natural transformation $\Phi : \mathscr{D} \dot{\to} \mathscr{U}$, which for each $M \in \mathrm{Obj}(\mathsf{Loc})$ is the canonical map

$$\Phi_M : \mathscr{D}(M) \to \mathscr{U}(M), \quad f \mapsto (0, f, 0, \dots),$$

is called the *(locally covariant) quantum field* associated to $\mathscr{U}$. Observe that every element in $\mathscr{U}(M)$ is a limit of sums and products of Φ_M applied to test functions because $\bigoplus_n \mathscr{D}(M)^{\otimes n}$ is dense in $\mathscr{U}(M)$.

---

[2]With this topology the algebra is complete and nuclear, and the algebra product is separately continuous.

5.3.2   Field equation and commutator

The Borchers–Uhlmann algebra carries no dynamical information and may therefore also be called the *off-shell field algebra*. In particular, the theory $\mathscr{U}$ is neither causal nor does it satisfy the timeslice axiom. To obtain a causal theory that satisfies the timeslice axiom, we need to implement a *field equation* (an *equation of motion*) that induces a Cauchy evolution of the algebra elements and a commutator that 'separates' causally disjoint algebra elements.

Therefore, we assign now to every globally hyperbolic spacetime M a natural, formally self-adjoint, Green-hyperbolic operator $P_M : \Omega^p(M,\mathbb{C}) \to \Omega^p(M,\mathbb{C})$ such that

$$\psi_* \circ P_M = P_N \circ \psi_*$$

for ever hyperbolic embedding $\psi : M \to N$. Moreover, we define a functor $\mathscr{D}$ as in the previous section.

As discussed in Sect. 3.6, associated to the Green-hyperbolic operator P_M, there exists on each globally hyperbolic manifold a unique causal propagator G_M.

The causal propagator is also called the commutator distribution or Pauli–Jordan distribution (see also Sect. 3.6.2) because it facilitates the definition of a *commutator* on the algebra $\mathscr{U}(M)$. Namely, let $f, h \in \mathscr{D}(M)$ then we define on $\mathscr{U}(M)$

$$[\Phi_M(f), \Phi_M(h)] \doteq (iG_M(f \otimes h), 0, \dots). \tag{5.1}$$

Due to the support properties of the causal propagator, this is exactly the right choice if one wants to implement Einstein causality.

The commutator extends to arbitrary elements of $\mathscr{U}(M)$ in the following way: First, we notice that the commutator ought to satisfy the Leibniz rule. Therefore it may be seen as a map

$$\mathscr{D}(M)^{\otimes n} \otimes \mathscr{D}(M)^{\otimes m} \longrightarrow \mathscr{D}(M)^{\otimes(n+m-2)}$$

for $n, m \geq 1$, which can be extended (anti-)linearly to $\bigoplus_n \mathscr{D}(M)^{\otimes n}$, a dense subalgebra of $\mathscr{U}(M)$. Finally, we can extend the resulting commutator continuously to the whole algebra $\mathscr{U}(M)$. Thereby the algebra becomes a Lie algebra.

The commutator is of immense physical importance. Foremost, it implements causality and manifests the principle of locality. Moreover, if the centre of the algebra of observables with respect to the commutator is non-trivial, the algebra contains unobservables and one cannot justify calling it 'algebra of observables'. Nevertheless, non-trivial centres in the 'algebra of observables' can lead to important non-local observable effects under spacetime embeddings [194].

Note that the commutator (5.1) defined on $\mathscr{U}(M)$ is degenerate if $\mathscr{D}(M) \cap \ker G_M$ is non-trivial and thus it leads to an algebra with a non-trivial centre. This problem will be addressed in the following section, where we introduce the so-called field algebra.

5.3.3   Field algebra

Then, taking the wave operator and the commutator, we can define the *(on shell) field algebra* $\mathscr{F}(M)$ as the unital topological *-algebra given for every $M \in \mathrm{Obj}(\mathsf{Loc})$

by the quotient

$$\mathscr{F}(M) \doteq \mathscr{U}(M)/\mathscr{I}(M),$$

where $\mathscr{I}(M)$ is the completion of the $*$-ideal finitely generated for all $f, h \in \mathscr{D}(M)$ by

(a) the wave equation

$$\Phi_M(\mathrm{P}_M f) \sim \Phi_M(0)$$

(b) the commutator relation

$$\Phi_M(f)\Phi_M(h) - \Phi_M(h)\Phi_M(f) \sim [\Phi_M(f), \Phi_M(h)]$$

The topology of $\mathscr{F}(M)$ is the quotient topology with respect to $\mathscr{U}(M)$.

Like for the Borchers–Uhlmann algebra, $\mathscr{F} : \mathsf{Loc} \to {}^*\mathsf{Alg}$ defines a functor, where $\mathscr{F}(M)$ is the field algebra and $\mathscr{F}(\psi)$ the $*$-algebra homomorphism

$$\mathscr{F}(\psi)[F] = [\psi_* F]$$

on all $[F] \in \mathscr{F}(M)$, which is naturally induced from $\mathscr{U}(\psi)$ via the canonical projection $[\,\cdot\,] : \mathscr{U}(M) \to \mathscr{F}(M)$. That these assignments give indeed a covariant functor, relies on the naturality of all involved operators. In particular,

$$\psi_*(\mathrm{P}_M f) = \mathrm{P}_N(\psi_* f) \quad \text{and} \quad G_M(f \otimes h) = G_N(\psi_* f \otimes \psi_* h)$$

for all $\psi : M \to N$ and $f, h \in \mathscr{D}(M)$. Note that the field algebra is a Lie algebra, where the bracket is simply

$$\big[[F], [H]\big] \doteq [FH - HF] = \big[[F, H]\big]$$

for all $[F], [H] \in \mathscr{F}(M)$; the centre is trivial by construction. Moreover, we can construct a quantum field ϕ for $\mathscr{F}$ as the natural transformation $\phi : \mathscr{D} \dashrightarrow \mathscr{F}$, which is given for each $M \in \mathsf{Obj}(\mathsf{Loc})$ by

$$\phi_M \doteq [\,\cdot\,] \circ \Phi_M.$$

That is, ϕ_M is the map $f \mapsto [(0, f, 0, \dots)]$ for all test functions $f \in \mathscr{D}(M)$.

The following is a standard result, see *e.g.* [190, Chap. 3.1]:

Proposition 5.1. *The locally covariant theory $\mathscr{F}$, given by the field algebra, satisfies both causality and the timeslice axiom.*

Proof. Since $\mathscr{F}(M)$ is the completion of the algebra generated by ϕ_M, causality follows immediately from the support properties of the causal propagator G_M.

For the timeslice axiom we only need to show that the algebra in the whole spacetime can be reconstructed from the algebra in a causally convex neighbourhood $N \subset M$ of a Cauchy surface; N may be considered as a spacetime in Loc with a

Cauchy embedding into M. Set $\chi \in C^\infty(M)$ such that $\chi = 1$ on $J^+(N) \setminus N$ and and $\chi = 0$ on $J^-(N) \setminus N$. For every $f \in \mathcal{D}(M)$ there exists $f' \in \mathcal{D}(M)$ given by

$$f' = P_M(\chi G_M f)$$

such that $\operatorname{supp} f' \subset N$ and

$$f - f' = P_M\big((1 - \chi)G_{\vee,M}f + \chi G_{\wedge,M}f\big) \in \ker G_M \cap \mathcal{D}(M)$$

The statement follows again because $\mathcal{F}(M)$ is the completion of the algebra generated by ϕ_M. $\square$

If ω is a state on the Borchers–Uhlmann algebra $\mathcal{U}(M)$ for some spacetime M, it induces a state for the field algebra $\mathcal{F}(M)$ if it also satisfies the commutation relation

$$\omega(FH) - \omega(HF) = \omega([F,H])$$

and the equation of motion

$$\omega\big((\mathrm{id} \otimes \cdots \otimes P_M \otimes \cdots \otimes \mathrm{id})F\big) = 0$$

for all $F, H \in \mathcal{U}(M)$. In that case, the state on $\mathcal{F}(M)$ is defined by the pushforward of ω by $[\,\cdot\,]$. Conversely, a state on $\mathcal{F}(M)$ always induces a state on $\mathcal{U}(M)$ by the pullback via $[\,\cdot\,]$.

5.3.4   Weyl algebra

The disadvantage of the Borchers–Uhlmann algebra and the field algebra is that they are only $*$-algebras and not C^*-algebras and hence cannot generally be represented by an algebra of bounded operators. However, heuristically speaking, we can exponentiate the field algebra to produce a C^*-algebra, the Weyl C^*-algebra introduced in Sect. 3.2.2.

In this section, set

$$\mathcal{D}(M) = \Omega_0^p(M, \mathbb{C})/\ker G_M.$$

$\mathcal{D}$ extends to a functor from Loc into the category of closed nuclear locally convex $\mathbb{C}$ vector spaces because G_M is continuous (so that $\ker G_M$ is closed) and transforms covariantly under hyperbolic embeddings (*cf.* the previous section).

Then define on every spacetime $M \in \mathsf{Loc}$ the Weyl C^*-algebra $\overline{\mathcal{W}}(M)$ obtained from the Weyl operators $W_M : \mathcal{D}(M) \to \overline{\mathcal{W}}(M)$ with commutator relation

$$W_M([f])W_M([h]) = \exp\big(\tfrac{\mathrm{i}}{2}G_M([f] \otimes [h])\big)W_M([f + h])$$

for every $[f], [h] \in \mathcal{D}(M)$ and where we denoted also by G_M the well-defined pushforward of G_M to $\mathcal{D}(M)$ via the canonical quotient map. Again, it may be shown that $\overline{\mathcal{W}}$ extends to a functor from Loc to *Alg by the covariance of the commutator distribution.[3] The proof that $\overline{\mathcal{W}}$ satisfies both causality and the timeslice axiom is very similar to Prop. 5.1 and will not be repeated.

---

[3]To be precise, it is also necessary to show that the minimal regular norm behaves in a locally covariant way. This follows from the Hahn–Banach theorem.

5.3.5 Scalar Klein–Gordon field

The (free, scalar) *Klein–Gordon equation* is

$$(\Box + \xi R + m^2)\varphi = 0, \tag{5.2}$$

where $\varphi \in \mathscr{E}(M)$ is the classical *Klein–Gordon field* and the parameters ξ and $m \geq 0$ are the *curvature coupling* and the *mass*.

One distinguishes in particular between two different curvature couplings: *minimal coupling* if $\xi = 0$ and *conformal coupling* if $\xi = 1/6$. The reason for naming $\xi = 1/6$ conformal coupling is that, in the massless case $m = 0$, (5.2) is invariant under conformal isometries, see *e.g.* [64, 216]. Namely, given a conformal embedding $\psi : M \to N$ with $\psi^* h = \Omega^2 g$, one finds

$$\psi^* \left(\Box + \tfrac{1}{6}R\right) \varphi = \Omega^3 \left(\Box + \tfrac{1}{6}R\right) \Omega^{-1} \psi^* \varphi,$$

where $\varphi \in \mathscr{E}(M)$. That is, if φ solves the massless conformally coupled Klein–Gordon equation on (N, h), then $\Omega^{-1}\psi^*\varphi$ solves the massless conformally coupled Klein–Gordon equation on (M, g).

The field algebra for the Klein–Gordon field can be constructed exactly as outlined above, where we choose

$$\mathrm{P}_M = \Box + \xi R + m^2 \quad \text{and} \quad \mathscr{D}_\varphi(M) = \mathscr{D}(M, \mathbb{C}),$$

and denote the resulting functor $\mathscr{F}_\varphi$ and the quantum field $\widehat{\varphi}$. The Weyl algebra may be constructed in a similar way. In case of conformal coupling, the Klein–Gordon field can also be quantized as a *conformally* locally covariant theory [177, 5].

5.3.6 Proca field

The field equation for the classical *Proca field* $Z \in \Omega^1(M)$ of mass $m > 0$ is

$$(\delta \mathrm{d} + m^2)Z = 0 \tag{5.3}$$

and one can almost immediately see that it is not normally hyperbolic. However, applying the codifferential to this equation we find that $\delta Z = 0$ so that (5.3) is equivalent to

$$(\Box + m^2)Z = 0$$

with the constraint $\delta Z = 0$.

There are two equivalent approaches to quantize this constrained system. Both rely on the same fact that the exterior derivative d and the codifferential δ commute with $(\Box + m^2)$ and thus also with its causal propagator G_M [94, 175].

For first approach [55, 94] we notice that (5.3) is pre-normally hyperbolic so that we can construct the causal propagator

$$\widetilde{\mathrm{G}}_M = \mathrm{G}_M \circ \left(m^{-2}\mathrm{d}\delta + \mathrm{id}\right).$$

Accordingly we can perform the construction of the field algebra discussed in Sect. 5.3.3 with this propagator. More precisely, we set

$$\mathrm{P}_M = \delta \mathrm{d} + m^2 \quad \text{and} \quad \mathscr{D}_Z(M) = \Omega_0^1(M, \mathbb{C})$$

and follow through with the construction of the field algebra, where we denote the resulting locally covariant theory $\mathscr{F}_Z$ and the corresponding quantum field $\widehat{Z}$.

The second possibility, which is related to the framework developed in [115], is to consider

$$\mathrm{P}_M = \square + m^2 \quad \text{and} \quad \mathscr{D}'_Z(M) = \{ f \in \Omega_0^1(M, \mathbb{C}) \mid \delta f = 0 \}$$

Since every section $f \in \mathscr{D}'_Z(M)$ is coclosed, $\mathrm{G}_M f$ solves the Proca equation (5.3). We can then perform the usual construction for the field algebra and denote the corresponding locally covariant theory $\mathscr{F}'_Z$.

Proposition 5.2. *The locally covariant theories $\mathscr{F}_Z$ and $\mathscr{F}'_Z$ are equivalent. That is, $\mathscr{F}_Z(M) \simeq \mathscr{F}'_Z(M)$ for every globally hyperbolic M.*

Proof. For every $f \in \mathscr{D}_Z(M)$

$$m^2 \widehat{Z}(f) = \widehat{Z}\big(m^2 f - (\delta \mathrm{d} + m^2)f\big) = -\widehat{Z}(\delta \mathrm{d} f)$$

and thus $\widehat{Z}\big(\mathscr{D}_Z(M)\big) = \widehat{Z}\big(\mathscr{D}'_Z(M)\big)$. Then it is easy to check that

$$(\delta \mathrm{d} + m^2)f = (\square + m^2)f \quad \text{and} \quad \widetilde{\mathrm{G}}_M f = \mathrm{G}_M f$$

for all $f \in \mathscr{D}'_Z(M)$.   □

Similarly, two apparently different Weyl algebras can be constructed for the Proca field and shown to be equivalent.

6

Quantum states

Summary

In Sect. 3.2.1 we already introduced some general features of states on *-algebras. In this section we will discuss features important or specific to quantum field theory.

We begin our discussion with the introduction of the n-point distributions (Sect. 6.1.1) associated to (some) states of the algebras defined above: the Borchers–Uhlmann algebra, the field algebra and the Weyl algebra. Of particular importance are states which satisfy the microlocal spectrum condition to be defined in Sect. 6.1.2. States which satisfy this constraint on the wavefront set are the so-called Hadamard states and their singular part is given by the Hadamard parametrix (Sect. 6.1.3).

After we introduced these general notions, we will discuss the construction of quantum states on particular spacetimes. Due to their importance in cosmology and their relative simplicity, we discuss adiabatic and Hadamard states on cosmological spacetimes in Sect. 6.2.

6.1 Preliminaries

Let (M, g) be a globally hyperbolic spacetime and let us consider, as in 5.3, p-form fields. It is important to notice that none of the results here are fundamentally restricted to the assumption of p-form field and can be easily generalized.

6.1.1 n-Point distributions

Let $\mathscr{U}(M)$ be the Borchers–Uhlmann algebra of a quantum field theory on M which is built on a test function space $\mathscr{D}(M) \subset \Omega_0^p(M)$. Since

$$\mathscr{U}(M) = \bigoplus_{n \in \mathbb{N}_0} \mathscr{D}(M)^{\widehat{\otimes} n},$$

the topological dual is of the form

$$\mathscr{U}'(M) = \prod_{n \in \mathbb{N}_0} \mathscr{D}'(M)^{\widehat{\otimes} n},$$

where we have used the kernel theorem. In other words, whereas any element of $\mathscr{U}(M)$ can be understood as a polynomial, $\mathscr{U}'(M)$ also contains power series. It follows that the any state ω on $\mathscr{U}(M)$ is uniquely defined by a family $(\omega_n)_{n \in \mathbb{N}}$ of *n-point distributions* (also called *n*-point functions or Wightman functions) $\omega_n \in \mathscr{D}'(M)^{\widehat{\otimes} n}$. If we denote by Φ_M the quantum field associated to $\mathscr{U}(M)$, then the n-point distributions satisfy

$$\omega_n(f_1 \otimes \cdots \otimes f_n) = \omega\big(\Phi_M(f_1) \cdots \Phi_M(f_n)\big)$$

for all $f_1,\ldots,f_n \in \mathscr{D}(M)$ and $n \in \mathbb{N}$.

The *connected* or *truncated n-point distributions* ω_n^T of a state ω are defined by the relation

$$\omega_n(x_1,\ldots,x_n) = \sum_{P \in \mathscr{P}_n} \prod_{r \in P} \omega_{|r|}^T(x_{r(1)},\ldots,x_{r(|r|)}), \tag{6.1}$$

where $\mathscr{P}_n$ denotes the (ordered) partitions of the set $\{1,\ldots,n\}$. Therefore, they can be calculated recursively from the n-point distributions. The first two truncated n-point distributions are

$$\omega_1^T(x_1) = \omega_1(x_1),$$
$$\omega_2^T(x_1,x_2) = \omega_2(x_1,x_2) - \omega_1(x_1)\omega_1(x_2)$$

and a general recursive formula is given by

$$\omega_n^T(x_1,\ldots,x_n) = \omega_n(x_1,\ldots,x_n) - \sum_{\substack{P \in \mathscr{P}_n \\ |P|>1}} \prod_{r \in P} \omega_{|r|}^T(x_{r(1)},\ldots,x_{r(|r|)}).$$

Thanks to the close relation of the Borchers–Uhlmann algebra and the field algebra, see the last paragraph of Sect. 5.3.3, the space of states of the field algebra is related to a subspace of the space of states for the Borchers–Uhlmann algebra and a state ω on the field algebra also has associated n-point distributions. These n-point distributions naturally satisfy the commutation relations

$$\omega_n(x_1,\ldots,x_n) = \omega_n(x_1,\ldots,x_{i+1},x_i,\ldots,x_n)$$
$$+ \omega_{n-2}(x_1,\ldots,\widehat{x}_i,\widehat{x}_{i+1},\ldots,x_n)G_M(x_i,x_{i+1}),$$

where the hats denote omitted points, and are weak solutions of the equations of motion

$$\mathrm{P}_M(x_i)\omega_n(x_1,\ldots,x_i,\ldots,x_n) = 0$$

for all $i \in [1..n]$. Therefore, if we denote by Φ_M the quantum field associated to $\mathscr{F}(M)$, the n-point distributions satisfy

$$\omega_n(f_1 \otimes \cdots \otimes f_n) = \omega\big(\Phi_M(f_1)\cdots\Phi_M(f_n)\big)$$

independently of the chosen representatives $f_1,\ldots,f_n \in \mathscr{D}(M)$ of $[f_1],\ldots,[f_n]$.

The definition of n-point distributions for a state on the Weyl algebra $\mathscr{W}(M)$ is slightly more involved. n-Point distributions in the algebraic sense only exist for strongly regular states as defined in Sect. 3.2.2, see also [23]. In this case they are defined by the relation

$$\omega_n(f_1 \otimes \cdots \otimes f_n) = (-\mathrm{i})^n \frac{\partial^n}{\partial t_1 \cdots \partial t_n} \omega\big(W_M(t_1[f_1])\cdots W_M(t_n[f_n])\big)\bigg|_{t_\bullet=0}.$$

Clearly, the n-point distributions of a state on the Weyl algebra satisfy the commutation relation and the equation of motion.

A state ω is called *quasi-free* or *Gaussian* if all its truncated n-point distributions vanish for $n \neq 2$, whence it is completely determined by its two-point distribution ω_2.

For a quasi-free state, all odd n-point distributions vanish and all even n-point distributions are given by

$$\omega_n(x_1,\ldots,x_n) = \sum_\sigma \omega_2(x_{\sigma(1)}, x_{\sigma(2)}) \cdots \omega_2(x_{\sigma(n-1)}, x_{\sigma(n)}),$$

where the sum is over all ordered pairings, *i.e.*, over all permutations $\sigma \in \mathfrak{S}_n$ such that $\sigma(1) < \sigma(3) < \cdots < \sigma(n-1)$ and $\sigma(1) < \sigma(2), \ldots, \sigma(n-1) < \sigma(n)$.

6.1.2   Microlocal spectrum condition

A quasi-free state ω satisfies the *microlocal spectrum condition* [45, 181, 189] if

$$\text{WF}(\omega_2) \subset \{(x, x'; \xi, -\xi') \in \dot{T}^*(M \times M) \,\big|\, (x; \xi) \sim (x'; \xi') \text{ and } \xi \rhd 0\}, \quad (6.2)$$

or, in words, the wavefront set of ω_2 is contained in the set of $(x, x'; \xi, -\xi') \in \dot{T}^*(M \times M)$ such that x, x' are connected by a lightlike geodesic γ with cotangent ξ at x and ξ' is the parallel transport of ξ to x' along γ (in symbols: $(x; \xi) \sim (x'; \xi')$) and ξ is future directed (in symbols: $\xi \rhd 0$). That is, $(x, x'; \xi, -\xi')$ is contained in the wavefront set if $(x; \xi)$ and $(x'; \xi')$ lie on the same future-directed bicharacteristic strip generated by $\sigma(\xi) = -g(\xi, \xi)$.

The microlocal spectrum condition can also be generalized to states that are not quasi-free [45, 191]. States that satisfy the microlocal spectrum condition are called *(generalized) Hadamard states*.

Let $P = \Box + B$, where B is a scalar function, the potential, and $G(x, x')$ its causal propagator. If the kernel of the two-point distribution ω_2 satisfies the commutator relation (weakly)

$$\omega_2(x, x') - \omega_2(x', x) = iG(x, x')$$

then equality of sets holds in (6.2). If, moreover, the two-point distribution is a parametrix of P_M, *i.e.* a weak bisolution up to smooth terms, then it attains the *local*[1] *Hadamard form* in a geodesically convex neighbourhood $U \subset M$

$$\begin{aligned}\omega_2(x, x') &= \lim_{\varepsilon \to 0^+} \frac{1}{8\pi^2} \left(\frac{u(x, x')}{\sigma_\varepsilon(x, x')} + v(x, x') \ln \frac{\sigma_\varepsilon(x, x')}{\lambda^2} + w(x, x') \right) \\ &= \mathscr{H}(x, x') + w(x, x'),\end{aligned} \quad (6.3)$$

where we take the *weak* limit, $x, x' \in U$, $\lambda \in \mathbb{R}$ is arbitrary and the detailed form of the coefficients $v, w \in \Gamma(\bigwedge^p(TM) \boxtimes \bigwedge^p(TM))$ will be discussed in the next section. Above we used a 'vectorized' van Vleck–Morette determinant

$$u(x, x') \doteq \Delta^{1/2}(x, x') g^{[a_1|b_1'}(x, x') \cdots g^{|a_p]b_p'}(x, x'),$$

which is antisymmetrized in the indices a_i and parallel transported along the geodesic connecting x and x', and the regularized world function

$$\sigma_\varepsilon(x, x') \doteq \sigma(x, x') + i\varepsilon \left(t(x) - t(x') \right) + \tfrac{1}{2}\varepsilon^2$$

with t a smooth time function on (M, g) compatible with the time-orientation.

---

[1]Also a *global* Hadamard form can be formulated [136], but since the discovery of the microlocal spectrum condition this global form has lost its importance. In fact, it was shown in [182] that a state that is everywhere locally of Hadamard form is also globally a Hadamard state.

6.1.3   The Hadamard parametrix

The coefficient functions

$$v(x,x') = \frac{1}{\lambda^2} \sum_{k=0}^{\infty} v_k(x,x') \left(\frac{\sigma(x,x')}{\lambda^2} \right)^k \tag{6.4}$$

in (6.3) are called *Hadamard coefficients* and are another example of bitensors. Note that the expansion above is an *asymptotic* expansion in terms of the world function σ and cannot be expected to converge unless the spacetime is analytic. Although it can be turned into a convergent series by replacing the series (6.4) by [24, Chap. 2.5]

$$\sum_{k=0}^{n} v_k(x,x') \sigma(x,x')^k + \sum_{k=n+1}^{\infty} v_k(x,x') \chi \left(\alpha_k^{-1} \sigma(x,x') \right)^k$$

for any $n \in \mathbb{N}$ and some sequence $(\alpha_k), \alpha_k \in (0,1]$, where $\chi \in C_0^{\infty}(\mathbb{R})$ is 1 in a neighbourhood of 0 (note that we omitted the scale λ), this will not concern us any further because we will only ever need a finite number of terms.[2] Therefore, we also define the truncated local Hadamard parametrix

$$\mathscr{H}_n(x,x') \doteq \lim_{\varepsilon \to 0^+} \frac{1}{8\pi^2} \left(\frac{u(x,x')}{\sigma_\varepsilon(x,x')} + \sum_{k=0}^{n} v_k(x,x') \sigma(x,x')^k \ln \frac{\sigma_\varepsilon(x,x')}{\lambda^2} \right).$$

One can show that there always exists a $n \in \mathbb{N}_0$ such that

$$\lim_{x' \to x} \mathscr{D} \left(\mathscr{H}(x,x') - \mathscr{H}_n(x,x') \right) = 0$$

for all differential operators $\mathscr{D}$ and n depends on the order of $\mathscr{D}$.

The coefficients v_k can be recursively calculated by (formally) applying P to $\mathscr{H}$; One then finds the so-called *Hadamard recursion relations* (*cf.* [69, 96, 169])

$$\lambda^2 Pu = (2\nabla_\sigma + \sigma^a{}_a - 2)v_0, \tag{6.5a}$$

$$\lambda^2 Pv_{k-1} = (2\nabla_\sigma + \sigma^a{}_a + 2k - 2)kv_k, \tag{6.5b}$$

where we have used the transport operators defined in (1.12). It can be shown that the Hadamard coefficients are symmetric in their arguments [103, 155]. Together with the first term in (6.3) the Hadamard coefficients make up the *Hadamard parametrix* $\mathscr{H}(x,x')$, which is therefore completely determined by the differential operator P and the geometry of the spacetime.

The covariant expansion of the Hadamard coefficients can be efficiently calculated using the Avramidi method described in Sect. 1.4.4 using the transport equations (6.5). If one is only interested in the coincidence limits, one can directly take the limit in (6.5) to find (omitting necessary Kronecker deltas originating from coincidence limits of parallel propagators)

$$[v_0] = \frac{1}{2}[Pu] = \frac{1}{2}\left(B - \frac{1}{6}R \right),$$

$$[v_k] = \frac{1}{2(1-k)k}[Pv_{k-1}], \quad k > 1.$$

---

[2]If we inserted this modification into (6.3), the two-point distribution would not be an exact (weak) solution of P any more, but only up to a smooth biscalar, *i.e.*, it would only be a parametrix.

Note that $[v_0]$ vanishes for a conformally coupled massless scalar field. The coincidence limit of v_1 cannot be found in this easy way and must be calculated directly. After a lengthy calculation using pencil and paper or a (fast) calculation using a tensor algebra software, one obtains (again omitting Kronecker deltas)

$$8[v_1] = B^2 + \tfrac{1}{3}\Box B - \tfrac{1}{3}RB + \tfrac{1}{36}R^2 - \tfrac{1}{90}R_{ab}R^{ab} + \tfrac{1}{90}R_{abcd}R^{abcd} - \tfrac{1}{15}\Box R. \qquad (6.6)$$

Different from $v(x,x')$, the symmetric bitensor $w(x,x')$ is not directly determined by the geometry or a differential operator. Instead the term $w(x,x')$ reflects the freedom in the choice of the state. Writing the asymptotic expansion

$$w(x,x') = \frac{1}{\lambda^2} \sum_{k=0}^{\infty} w_k(x,x') \left(\frac{\sigma(x,x')}{\lambda^2} \right)^k,$$

we notice that the freedom to choose a state is completely encoded in the first coefficient w_0 and the remaining coefficients obey the recursion relation [69, 96]

$$\lambda^2 P w_k = 2(k+1)\nabla_\sigma w_{k+1} + 2k(k+1)w_{k+1} + (k+1)w_{k+1}\sigma^a{}_a$$
$$+ 2\nabla_\sigma v_{k+1} - 2(2k+1)v_{k+1} + v_{k+1}\sigma^a{}_a.$$

A common choice is to set $w_0 = 0$ as in [219]. In any case, w_0 must be chosen such that w is symmetric.

6.2 Construction of states on cosmological spacetimes

Explicit examples of quantum states are known only on a small class of highly symmetric spacetimes. Below we will first discuss the so-called Bunch–Davies state [10, 51, 198], which can be considered the vacuum state of de Sitter spacetime. Then we study a construction of states on FLRW spacetimes due to Parker [172] called adiabatic states. Although adiabatic states are in general not Hadamard, indeed only adiabatic states of infinite order satisfy the microlocal spectrum condition [134], they can be considered approximate Hadamard states and have proven to be very useful thanks to their relatively straightforward construction. Since we will make extensive use of adiabatic states when we discuss the semiclassical Einstein equation on cosmological backgrounds in Chap. 8, they will be treated in some detail below. Moreover, we will introduce the states of low energy by Olbermann [165], which are constructed via a careful Bogoliubov transformation from adiabatic states.

6.2.1 Bunch–Davies state

A distinguished Hadamard state for the (massive) scalar field on de Sitter spacetime is the *Bunch–Davies state* [10, 51, 198]. It is the unique pure, quasi-free Hadamard state invariant under the symmetries of de Sitter spacetime. Note that equations of motion for the scalar field on de Sitter spacetime are

$$\Box\varphi + (12\xi H^2 + m^2)\varphi = 0,$$

where H is the Hubble constant, m the mass of the scalar field and ξ the curvature coupling, *cf.* Sect. 5.3.5. Therefore the curvature coupling ξ acts like a mass and

we set $M^2 = 12\xi H^2 + m^2$. The Bunch–Davies state is also a Hadamard state in the limit $M = 0$ but in that case it fails to be invariant under the symmetries of de Sitter spacetime [10, 11]; below we assume $M > 0$.

Using the function Z defined in (2.15), the Bunch–Davies state is the quasifree state given by the two-point distribution[3]

$$\omega_2(x, x') \doteq \omega_2(Z(x, x')) \doteq \frac{M^2 - 2H^2}{8\pi \cos(\pi v)} \, {}_2F_1\big(v_+, v_-; 2; \tfrac{1}{2}(1 + Z)\big), \qquad (6.7)$$

where ${}_2F_1$ is the analytically continued hypergeometric function and with

$$v_\pm \doteq \frac{3}{2} \pm v \quad \text{and} \quad v \doteq \sqrt{\frac{9}{4} - \frac{M^2}{H^2}}.$$

We can rewrite (6.7) into a form which exhibits the Hadamard nature of the state more clearly. In fact, using well-known transforms [166, Eq. (15.8.10)] of the hypergeometric function ${}_2F_1$, one can show

$$\omega_2(Z) = \frac{H^2}{8\pi^2}(1 - Z)^{-1} + \frac{M^2 - 2H^2}{8\pi^2}\Big(\widetilde{v}\big(\tfrac{1}{2}(1 - Z)\big) \ln\big(\tfrac{1}{2}(1 - Z)\big) + \widetilde{w}\big(\tfrac{1}{2}(1 - Z)\big)\Big)$$

for $|Z| < 1$ (spacelike separated points) with

$$\widetilde{v}(z) = {}_2F_1(v_+, v_-; 2; z),$$

$$\widetilde{w}(z) = \sum_{k=0}^{\infty} \frac{(v_+)_k (v_-)_k}{k!(k+1)!} \big(F(v_+ + k) + F(v_- + k) - F(k+1) - F(k+2)\big) z^k,$$

where F is the digamma-function.

In the cosmological chart of de Sitter spacetime the function Z attains the simple form (2.17) and the spatial sections are flat. Therefore, one can represent ω_2 as a spatial Fourier transformation with respect to $\vec{x} - \vec{x}'$. Indeed, using known integrals of (modified) Bessel functions [107, §6.672], a lengthy calculation shows [198] (omitting again the ε-prescription)

$$\omega_2(x, x') = \frac{H^2(\tau\tau')^{3/2}}{32\pi^2} \int_{\mathbb{R}^3} e^{-\pi \operatorname{Im} v} H_v^{(1)}(-k\tau) H_{\bar{v}}^{(2)}(-k\tau') e^{i\vec{k}\cdot(\vec{x} - \vec{x}')} \, d\vec{k}, \qquad (6.8)$$

where $x = (\tau, \vec{x})$ and $x' = (\tau', \vec{x}')$ in the conformal coordinates and $H^{(1)}, H^{(2)}$ are the Hankel functions of first and second kind.

6.2.2   Homogeneous and isotropic states

It is usually reasonable to restrict ones attention to states that respect the symmetry of the background spacetime. Under this assumption, a state on a FLRW spacetime should be both homogeneous and isotropic. That is, if the state is also quasifree, its two-point distribution needs to satisfy

$$\omega_2(x, x') = \omega_2(t, t', \vec{x} - \vec{x}') = \omega_2(\tau, \tau', \vec{x} - \vec{x}'),$$

---

[3]More precisely, one should replace Z by $Z + i\varepsilon(t(x) - t(x'))$, where t is a smooth time function, and take the limit $\varepsilon \to 0^+$. Note that ${}_2F_1$ has a branch cut from 1 to ∞.

where $x = (t, \vec{x}) = (\tau, \vec{x})$ and $x' = (t', \vec{x}') = (\tau', \vec{x}')$ with respect to cosmological or conformal time.

Under a certain relatively weak continuity assumption on the two-point distribution (such that they may be represented as bounded operators on a certain Hilbert space and the Riesz representation theorem can be used [146]), it was shown in [146, 196] that every *quasifree, homogeneous and isotropic state* for the scalar field is of the form[4]

$$\omega_2(x, x') = \frac{1}{(2\pi)^3 a(\tau)a(\tau')} \int_{\mathbb{R}^3} \big(\Xi(k)S_k(\tau)\overline{S}_k(\tau') \tag{6.9}$$
$$+ (\Xi(k) + 1)\overline{S}_k(\tau)S_k(\tau') \big)\, e^{i\vec{k}\cdot(\vec{x}-\vec{x}')}\, d\vec{k},$$

where $k = |\vec{k}|$ and $\Xi(k)$ is a non-negative (almost everywhere), polynomially bounded function in $L^1(\mathbb{R}_0^+)$; for pure states $\Xi = 0$. Moreover, the modes S_k are required to satisfy the *mode equation* of motion

$$(\partial_\tau^2 + \omega_k^2)S_k(\tau) = 0, \quad \omega_k^2 \doteq k^2 + (\xi - \tfrac{1}{6})a(\tau)^2 R + a(\tau)^2 m^2, \tag{6.10}$$

and the Wronski-determinant condition[5]

$$\overline{S}_k S_k' - \overline{S}_k' S_k = i, \tag{6.11}$$

where both S_k and S_k' are polynomially bounded in k. States constructed in this manner are in general *not* of Hadamard type.

Two important examples of pure Hadamard states expressible in the mode form above are the *Minkowski vacuum state* on Minkowski spacetime

$$\frac{1}{(2\pi)^3} \int_{\mathbb{R}^3} \frac{1}{2E(k)} e^{-iE(k)(t-t')} e^{i\vec{k}\cdot(\vec{x}-\vec{x}')}\, d\vec{k},$$

with $E(k)^2 = k^2 + m^2$, and the Bunch–Davies state on the cosmological patch of de Sitter spacetime (6.8). Interesting examples of non-pure states are the *approximate KMS states* at inverse temperature β for the conformally coupled scalar field [57]

$$\frac{1}{(2\pi)^3 a(\tau)a(\tau')} \int_{\mathbb{R}^3} \left(\frac{S_k(\tau)\overline{S}_k(\tau')}{e^{\beta k_F} - 1} + \frac{\overline{S}_k(\tau)S_k(\tau')}{1 - e^{-\beta k_F}} \right) e^{i\vec{k}\cdot(\vec{x}-\vec{x}')}\, d\vec{k}$$

with $k_F = \sqrt{k^2 + a(\tau_F)^2 m^2}$ for some 'freeze-out' time τ_F. These states are KMS states if the spacetime admits a global timelike Killing vector field which is a symmetry of the state; they are Hadamard states if the pure state specified by the modes S_k is already a Hadamard state [57].

Given fixed reference modes χ_k that satisfy (6.10) and (6.11), all other possible mode solutions S_k can be constructed via a *Bogoliubov transformation*, i.e.,

$$S_k = A(k)\chi_k + B(k)\overline{\chi}_k \quad \text{with} \quad |A(k)|^2 - |B(k)|^2 = 1,$$

---

[4]Note that we omit here and below the necessary ε-regularization of the integral, where one multiplies the integrand with $e^{-\varepsilon k}$ and considers the weak limit $\varepsilon \to 0^+$.

[5]Imposing the Wronski-determinant condition guarantees that the imaginary part of the two-point distribution is given by half the commutator distribution. It is sufficient to impose this condition at one instance in time.

where $A(k)$ and $B(k)$ are such that $k \mapsto S_k$ and $k \mapsto S'_k$ are (essentially) polynomially bounded, measurable functions. Note that changing S_k by a phase does not affect the state and therefore $B(k)$ can always be chosen to be real. The choice of A and B thus corresponds to two degrees of freedom, *e.g.*, the phase of A and the modulus of B. If the reference modes χ_k specify a pure Hadamard state, one can show that the modes S_k with the mixing $\Xi(k)$ specify a Hadamard state as well if and only if (in addition to the conditions above) $k^n B(k)$ and $k^n \Xi(k)$ are in $L^1(\mathbb{R}_0^+)$ for all $n \in \mathbb{N}$ and $\operatorname{Arg} A - \operatorname{Arg} B$ is measurable [178, 221]. An important example of a Bogoliubov transformation of a Hadamard state that (clearly) does *not* give a Hadamard state are the α-vacua associated to the Bunch–Davies state for which $A = \sinh \alpha$ and $B = \cosh \alpha$ with $\alpha \in \mathbb{R}$.

6.2.3   Adiabatic states

Any solution of (6.10) and (6.11) is of the form[6]

$$S_k(\tau) = \frac{\rho_k(\tau)}{\sqrt{2}} \, e^{i\theta_k(\tau)} \quad \text{with} \quad \theta_k(\tau) = \int^\tau \rho_k(\eta)^{-2} \, d\eta, \tag{6.12}$$

where ρ_k satisfies the differential equation

$$\rho_k'' = (\rho_k^{-4} - \omega_k^2)\rho_k = (\theta_k'^2 - \omega_k^2)\rho_k. \tag{6.13}$$

Modes $W_k = \sigma_k e^{i\psi_k}/\sqrt{2}$ of the form (6.12) with arbitrary σ_k that do *not* satisfy the differential equation (6.13) can be used to specify initial values for solutions S_k of the mode equation (6.10), *i.e.*,

$$S_k(\tau_0) = W_k(\tau_0), \quad S'_k(\tau_0) = W'_k(\tau_0)$$

at some initial time τ_0. Using an idea of Parker [172], it can be shown that the W_k yield the solution S_k via a Bogoliubov-like transformation

$$S_k(\tau) = A(\tau)W_k(\tau) + B(\tau)\overline{W}_k(\tau) \tag{6.14}$$

with time-dependent coefficients given by

$$A(\tau) = 1 - i \int_{\tau_0}^\tau G(\eta) \left(A(\eta) + B(\eta) e^{-2i\psi_k(\eta)} \right) d\eta, \tag{6.15a}$$

$$B(\tau) = - \int_{\tau_0}^\tau A'(\eta) e^{2i\psi_k(\eta)} \, d\eta \tag{6.15b}$$

and $2G \doteq \sigma_k^{-2} - \omega_k^2 \sigma^2 - \sigma_k \sigma_k''$, where we have suppressed the k-dependence of A, B and G in all four equations above. Applying arguments from the analysis of Volterra integrals, it can be shown that $1 - A$, A', B and B' have the same large k behaviour as G (*cf.* [146, 165]). As a consequence, the modes S_k and their derivatives have (almost) the same asymptotic behaviour as the modes W_k if σ_k looks asymptotically like $\omega_k^{-1/2}$:

---

[6]The lower limit in the integration is arbitrary as it gives a constant phase.

Proposition 6.1. *Suppose* $\partial_\tau^l \sigma_k = \partial_\tau^l(\omega_k^{-1/2}) + O(k^{-9/2})$ *for all* $l = 0, 1, 2$ *such that* $G = O(k^{-m})$ *for some* $m \geq 3$. *The modulus* ρ_k *of the modes* S_k *satisfies*

$$\rho_k = \sigma_k + O(k^{-1/2-m}), \quad \partial_\tau^n \rho_k = \partial_\tau^n \sigma_k + O(k^{3/2-m})$$

for all $n = 0, 1, 2$.

Proof. First note that the assumed bounds of σ_k and σ_k'' yield

$$2G = \sigma_k^{-2} - \omega_k^2 \sigma^2 - \sigma_k \sigma_k'' = O(k^{-3}),$$

which is consistent with the assumption on G. Further, recall that $1 - A$, A', B and B' are $O(k^{-m})$, too. We then derive from (6.14) that

$$\rho_k^2 = |A + B\,e^{-2i\psi_k}|^2 \sigma_k^2 = \sigma_k^2 + O(k^{-1-m})$$

and thus $\rho_k = \sigma_k + O(k^{-1/2-n})$. Next we use that ρ_k satisfies the differential equation (6.13) to find

$$\rho_k'' - \sigma_k'' = (\rho_k^{-4} - \omega_k^2)\rho_k - (\sigma_k^{-4} - \omega_k^2)\sigma_k + 2G\sigma_k^{-1} = O(k^{3/2-m}),$$

whereby we obtain the estimate for ρ_k'' and, after an integration in time, also that for ρ_k'. $\qquad \square$

In summary, the asymptotic behaviour of the initial values given by ω_k fixes the asymptotic behaviour of the solutions ω_k.

We can now construct *adiabatic states* as in [146, 172] by specifying appropriate initial values for (6.10) respectively (6.13): Making a WKB-like Ansatz, one finds the adiabatic modes of Parker [172]. Namely, the adiabatic modes $W_k = W_k^{(n)}$ of order n are modes of the form (6.12) with $\sigma_k = \sigma_k^{(n)}$, given iteratively via[7]

$$(\sigma_k^{(n+1)})^{-4} \doteq \omega_k^2 + \frac{\sigma_k^{(n)''}}{\sigma_k^{(n)}} \quad \text{with} \quad (\sigma_k^{(0)})^{-4} \doteq \omega_k^2.$$

The adiabatic modes $W_k^{(n)}$ are then used to specify initial values for the mode equation (6.10), *e.g.*, by solving the integral equations (6.15).

A useful result on the asymptotic behaviour of the adiabatic modes is stated in [146, Lem. 3.2]. Using the fact that $\omega_k' = O(k^{-1})$, one can easily improve this lemma to obtain for all $n \in \mathbb{N}_0$ and $m \in \mathbb{N}$

$$\begin{aligned}
\sigma_k^{(n)} &= O(k^{-1/2}), & \varepsilon_k^{(n)} &= O(k^{-2(n+1)}), \\
\frac{\partial^m}{\partial\tau^m}\sigma_k^{(n)} &= O(k^{-5/2}), & \frac{\partial^m}{\partial\tau^m}\varepsilon_k^{(n)} &= O(k^{-2(n+1)})
\end{aligned} \tag{6.16}$$

as $k \to \infty$ and where $(\sigma_k^{(n)})^{-4} = (\sigma_k^{(n-1)})^{-4}(1 + \varepsilon_k^{(n)})$. The asymptotic behaviour of the coefficients (6.15) for adiabatic states was analyzed in [146, 165]. It can be found, using the improved bounds (6.16), that they satisfy

$$\begin{aligned}
1 - A(k, \tau) &= O(k^{-2n-3}), & B(k, \tau) &= O(k^{-2n-3}), \\
A'(k, \tau) &= O(k^{-2n-3}), & B'(k, \tau) &= O(k^{-2n-3})
\end{aligned}$$

---

[7]The notation used here can be transformed into the usual one by setting $\sigma_k^{(n)} = (\Omega_k^{(n)})^{-1/2}$.

as $k \to \infty$.

These results can be seen as a starting point to show the relation between adiabatic states of a certain order and Hadamard states. Indeed, one can show [134] that adiabatic states of infinite order are Hadamard states and that adiabatic states of a finite order satisfy a microlocal spectrum condition on a certain *Sobolev wavefront set*[8] of the two-point distribution.

6.2.4   An adiabatic state for conformal coupling

Let study the construction of the (adiabatic) states already considered in [178, 3], see also [14, 15]. For a (massive) conformally coupled scalar field ($\xi = 1/6$) the initial values for an adiabatic state of order zero can be taken to be

$$\chi_k(\tau_0) = \frac{1}{\sqrt{2k_0}} e^{ik_0 \tau_0}, \qquad \chi_k'(\tau_0) = \frac{ik_0}{\sqrt{2k_0}} e^{ik_0 \tau_0},$$

with $k_0^2 \doteq (\Omega_k^{(0)})^2 = k^2 + a(\tau_0)^2 m^2$. Note that these initial values are essentially a conformal transformation of the modes of the Minkowski vacuum.

It is possible solve the mode equation (6.10) with these initial values perturbatively. For this purpose, define the potential $V(\tau) = m^2(a(\tau)^2 - a(\tau_0)^2)$ and make the recursive Ansatz $\chi_k(\tau) = \sum_n \chi_k^{(n)}(\tau)$ with the recurrence relation

$$\chi_k^{(n)\prime\prime}(\tau) + k_0^2 \chi_k^{(n)}(\tau) = -V(\tau)\chi_k^{(n-1)}(\tau) \tag{6.17}$$

with initial condition

$$\chi_k^{(0)}(\tau) \doteq \frac{1}{\sqrt{2k_0}} e^{ik_0 \tau}. \tag{6.18}$$

The mode equation (6.10) is then solved as described in the proof of the following proposition:

Proposition 6.2. *The recurrence relation (6.17) is solved iteratively (for $\tau > \tau_0$) by*

$$\chi_k^{(n)}(\tau) = \int_{\tau_0}^{\tau} \frac{\sin\left(k_0(\eta - \tau)\right)}{k_0} V(\eta)\chi_k^{(n-1)}(\eta)\, d\eta \tag{6.19}$$

and the sum $\chi_k(\tau) = \sum_n \chi_k^{(n)}(\tau)$ converges.

Proof. Consider for each k the retarded propagator of $\partial_\tau^2 + k_0^2$ given by

$$\Delta_{\mathrm{ret},k}(f)(\tau_0, \tau) = \int_{\tau_0}^{\tau} \frac{\sin\left(k_0(\tau - \eta)\right)}{k_0} f(\eta)\, d\eta, \quad \tau > \tau_0,$$

for all $f \in C^0(I)$, where I is the domain of the conformal time. Applying $\Delta_{\mathrm{ret},k}$ to (6.17), it can be solved as

$$\chi_k^{(n)}(\tau) = \int_{\tau_0}^{\tau} \frac{\sin\left(k_0(\eta - \tau)\right)}{k_0} V(\eta)\chi_k^{(n-1)}(\eta)\, d\eta$$

---

[8]Sobolev wavefront sets are very similar to the usual wavefront set, but instead of using smooth functions at the foundation of the definition, one uses functions from a Sobolev set of a certain order.

for $\tau > \tau_0$.

The straightforward estimates

$$|\chi_k^{(n)}| \le \frac{m^2}{k_0} \int_{\tau_0}^{\tau} \left|V(\eta)\chi_k^{(n-1)}(\eta)\right| \mathrm{d}\eta, \quad |\chi_k^{(n)}| \le m^2 \int_{\tau_0}^{\tau} (\tau - \eta)\left|V(\eta)\chi_k^{(n-1)}(\eta)\right| \mathrm{d}\eta$$

can be iterated using the initial value $\chi_k^{(0)} = (2k_0)^{-1/2}$ and the standard 'trick' of extending the integration of a symmetric function over a time-ordered domain to an integration over a symmetric domain by diving through the appropriate factorial (*cf.* [178, Prop. 4.4]). Combining the two estimates, this gives

$$|\chi_k^{(n)}| \le \frac{1}{\sqrt{2k_0}\, n!} \left(\frac{m^2}{k_0} \int_{\tau_0}^{\tau} \left|V(\eta)\right| \mathrm{d}\eta \right)^l \left(m^2 \int_{\tau_0}^{\tau} (\tau - \eta)\left|V(\eta)\right| \mathrm{d}\eta \right)^{n-l} \tag{6.20}$$

for any $0 \le l \le n$. Therefore the sum $\chi_k(\tau) = \sum_n \chi_k^{(n)}(\tau)$ converges absolutely. □

Equivalent results can be found in [3, Sect. 2.1], [178, Prop. 4.4] and also [14]. It is clear, that the recurrence relation can be solved in the same for $\tau < \tau_0$ by applying the advanced propagator.

Remark 6.3. *The partial modes $\chi_k^{(n)}$ can be computed as in Prop. 6.2 even if the metric (equivalently the scale factor) is not smooth. If the scale factor is C^0, the resulting mode χ_k will be at least C^2. This relies crucially on the fact that the curvature, which is not well-defined if a is not at least C^2, does not enter the mode equation (6.10).*

6.2.5   States of low energy

Let us *define* the *(unregularized) energy density per mode S_k* as

$$\hat{\rho}(S_k, \overline{S}_k) \doteq \frac{1}{2a^4} \Big(S_k' \overline{S}_k' + (6\xi - 1)aH(S_k \overline{S}_k)' \tag{6.21}$$
$$+ (k^2 + a^2m^2 - (6\xi - 1)a^2H^2)S_k \overline{S}_k \Big).$$

For now, we will not interpret this quantity in any way and leave its derivation to Sect. 7.3.

Given reference modes χ_k and Bogoliubov coefficients A, B, the energy density per mode $S_k = A\chi_k + B\overline{\chi}_k$ is related to the energy density per reference mode χ_k by

$$\tfrac{1}{2}\big(\hat{\rho}(S_k, \overline{S}_k) - \hat{\rho}(\chi_k, \overline{\chi}_k)\big) = |B|^2 \hat{\rho}(\chi_k, \overline{\chi}_k) + \mathrm{Re}\left(A\overline{B}\,\hat{\rho}(\chi_k, \chi_k)\right) \tag{6.22}$$

One can now attempt to minimize the energy density per mode by varying the Bogoliubov coefficients and we find that:

Proposition 6.4. *The energy density per mode at a fixed instance of time is minimal if and only if the Bogoliubov coefficients are given by (up to unitary equivalence)[9]*

$$\mathrm{Arg}\, A(k) = \pi - \mathrm{Arg}\left(\hat{\rho}(\chi_k, \chi_k)\right), \tag{6.23a}$$

$$B(k) = \left(\frac{\hat{\rho}(\chi_k, \overline{\chi}_k)}{2\sqrt{\hat{\rho}(\chi_k, \overline{\chi}_k)^2 - |\hat{\rho}(\chi_k, \chi_k)|^2}} - \frac{1}{2} \right)^{1/2} \tag{6.23b}$$

---

[9]Recall that, without loss of generality, we can always choose B to be real and positive such that A is completely determined by its phase.

and $\hat{\rho}(\chi_k, \overline{\chi}_k)^2 > |\hat{\rho}(\chi_k, \chi_k)|^2$. *The inequality is satisfied for all $k > 0$ if and only if*

$$\sqrt{1 + \frac{4m^2}{H^2}} \pm 1 \geq \pm 12\xi \qquad \text{or} \qquad H = 0, \tag{6.24}$$

i.e., in particular whenever $0 \leq \xi \leq 1/6$.

Proof. For fixed $B > 0$, (6.22) is minimized by $\text{Arg}\, A = \pi - \text{Arg}\left(\hat{\rho}(\chi_k, \chi_k)\right)$ so that the second summand is maximally negative. Therefore, minimizing (6.22) is equivalent to finding the minima of

$$B^2 \hat{\rho}(\chi_k, \overline{\chi}_k) + \text{Re}\left(AB\hat{\rho}(\chi_k, \chi_k)\right) = B^2 \hat{\rho}(\chi_k, \overline{\chi}_k) - \sqrt{1 + B^2}\, B|\hat{\rho}(\chi_k, \chi_k)|.$$

Differentiating this expression by B, one finds that an extremum exists on the positive real axis only if $\hat{\rho}(\chi_k, \overline{\chi}_k)^2 > |\hat{\rho}(\chi_k, \chi_k)|^2$ and that its locus is given by (6.23b); it is easy to see that this is indeed an minimum.

Inserting the definition (6.21) into the condition $\hat{\rho}(\chi_k, \overline{\chi}_k)^2 > |\hat{\rho}(\chi_k, \chi_k)|^2$, we find that it is equivalent to

$$k^2 + a^2\left(m^2 + 6(1 - 6\xi)\xi H^2\right) > 0.$$

If this conditions is to hold for all $k > 0$, then (6.24) must be satisfied. $\qquad\square$

Instead of trying to minimize the energy density per mode at an instant, states of low energy are constructed by minimizing the smeared energy density per mode. That is, by minimizing

$$\frac{1}{2} \int_I f(\tau)^2 \left(\hat{\rho}(S_k, \overline{S}_k) - \hat{\rho}(\chi_k, \overline{\chi}_k)\right) d\tau$$
$$= \int_I f(\tau)^2 \left(|B|^2 \hat{\rho}(\chi_k, \overline{\chi}_k) + \text{Re}\left(A\overline{B}\,\hat{\rho}(\chi_k, \chi_k)\right)\right) d\tau \tag{6.25}$$

for a fixed smearing function $f \in C_0^\infty(I)$, where $I \subset \mathbb{R}$ is the domain of the conformal time coordinate.

This minimization was performed for minimally coupled scalar fields in [165] to find the so-called *states of low energy*. It can be shown that the states of low energy satisfy the microlocal spectrum condition and thus they are Hadamard states. The arguments presented in [165] can be straightforwardly repeated for the conformally coupled scalar field to find states of low energy, which are Hadamard states too. In both cases the Bogoliubov coefficients are given as in (6.23) with the replacements

$$\hat{\rho}(\chi_k, \overline{\chi}_k) \to \int_I f(\tau)^2 \hat{\rho}(\chi_k, \overline{\chi}_k)\, d\tau \quad \text{and} \quad \hat{\rho}(\chi_k, \chi_k) \to \int_I f(\tau)^2 \hat{\rho}(\chi_k, \chi_k)\, d\tau.$$

There are good reasons to believe the following:

Conjecture 6.5. *States of low energy for arbitrary smearing function, mass and scale factor exist only in the curvature coupling range $0 \leq \xi \leq 1/6$. For all such ξ the state satisfies the microlocal spectrum condition.*

The crucial point in proving this conjecture is to show that (6.25) has a minimum. This can be shown similar to [165] for the minimally coupled case and the conformally coupled case. For other values of ξ the proof is more difficult, but by continuity it is clear from Prop. 6.4 that for some smearing functions states of low energy exist in the interval $\xi \in [0, 1/6]$ but not outside that range. Namely, if (f_n) is a sequence of functions such that f_n^2 converges weakly to the delta distribution, then there exists N such that a state of low energy exists for all f_m, $m \geq N$ because a minimum exists for $f^2 = \delta$ by Prop. 6.4. Once existence is show, one can expect that the state satisfies the microlocal spectrum condition using proofs analogous to those in [165].

6.3   Holographic construction of Hadamard states

In the absence of a global timelike Killing field on a generic globally hyperbolic spacetime it is difficult to find physically well-motivated quantum states. Therefore, in recent years, a lot of effort was put into the construction of proper Hadamard states on non-trivial spacetimes. A promising method is the 'holographic' construction of Hadamard states on characteristic surfaces introduced in [64, 157, 159]. The holographic method has been applied to construct Hadamard states for the conformally coupled, massless scalar field [64, 157, 159], the Weyl (massless Dirac) field [58, 112], the vector potential [1, 5] and linearized gravity [29] on asymptotically flat spacetimes and cosmological backgrounds [61, 62]. It was also used to construct local Hadamard states in lightcones in [65].

Forgetting for a moment the application of the bulk-to-boundary construction to asymptotically flat spacetimes and limiting ourselves to the scalar field, it may be roughly sketched as follows (see also [106]). Let (M, g) be a globally manifold with a distinguished point p such that the future lightcone of p satisfies some technical conditions. The it is possible to construct on the lightcone (without the tip and as a manifold on its own) a positive bidistribution λ on all functions on the lightcone that are compactly supported to the future and falls off sufficiently fast to the past, such that the antisymmetric part of λ agrees with the pullback of the commutator distribution on the whole spacetime, and the wavefront set of λ is of positive frequency with respect to the future-directed lightlike geodesics through p. This bidistribution has all the necessary properties to define a state for a quantum field theory on the lightcone. Moreover, taking any compactly supported function in the interior of the lightcone of p, it can be mapped to a function on the lightcone using the advanced propagator and a pullback such that the resulting function on the lightcone is compact towards the future and has good fall-off properties towards the past of the lightcone. This way one obtains the so-called bulk-to-boundary (projection) map. Pulling back all functions in the interior of the lightcone to the boundary of the lightcone using this map, one thus finds a state for the scalar field restricted to the interior of the lightcone. Applying the propagation of singularities theorem it is possible to show that the resulting state satisfies the microlocal spectrum condition.

In a second step, on may construct Hadamard states for conformally invariant scalar field on asymptotically flat spacetimes with globally hyperbolic unphysical spacetimes. First, one notices that boundary of the conformal completion of an asymptotically flat spacetime in the unphysical spacetimes satisfies all the necessary

technical conditions. Then one can compose the bulk-to-boundary map inside the unphysical spacetime with the (non-unique) conformal transformation associated with the asymptotically flat spacetime, to find a state for the conformally invariant scalar field. Since conformal transformation leave lightlike geodesics invariant, also this state is of Hadamard form.

In [1, 5] this construction was generalized to the electromagnetic vector potential. Also in this more complicated case a bulk-to-boundary construction of Hadamard states was be found, but it involves careful use of the gauge freedom of the vector potential to construct a *positive* state. Otherwise, the construction remains largely unchanged.

III

Semiclassical gravity

It is shown in a quite general manner that the quantization of a given system implies also the quantization of any other system to which it can be coupled. — Bryce S. DeWitt,
in "Gravitation: An Introduction to Current Research" (1962), p. 272

Oh gravity, thou art a heartless bitch.
— Sheldon Cooper, Season 1, Episode 2, The Big Bang Theory

7

The semiclassical Einstein equation

7.1 Introduction

The equation

$$G_{ab} + \Lambda g_{ab} = \omega(:T_{ab}:) \tag{7.1}$$

is called the *semiclassical Einstein equation*.[1] It is obtained from the ordinary Einstein equation by replacing the classical stress-energy tensor with the (normal ordered) expectation value of the stress-energy tensor of a quantum field in a suitable quantum state ω. Many developments in quantum field theory on curved spacetimes were driven by problems related to the quantum stress-energy tensor. See also the monographs [36, 104, 112, 217] for an overview of the subject.

The semiclassical Einstein equation is usually understood as an equation that describes physics midway between the classical regime covered by the Einstein equation (2.9) and a full-fledged, but still elusive, quantum gravity. Namely, in the semiclassical Einstein equation one takes into account that the 'matter' content of the universe is fundamentally of quantum nature as described by quantum field theory on curved spacetimes, whereas the background structure which is the spacetime is treated on a classical level and is not separately quantized.

On the right-hand side one usually considers only Hadamard states. The reason for restricting to Hadamard states is that only for Hadamard states the higher moments

$$\omega(:T_{ab}(x)::T_{ab}(x):) \quad etc.$$

can be defined. This is due to the fact that the n-point distributions of a state are distributions and thus they cannot simply be multiplied (*cf.* Sect. 3.5.7). Since the two-point distributions of Hadamard states satisfy the microlocal spectrum condition, their wavefront set is contained inside a convex cone in $\dot{T}^*(M \times M)$ and hence powers of the n-point distribution are well-defined distributions. This will be discussed in more detail in Chap. 9.

7.2 The stress-energy tensor

While the left-hand side of the semiclassical Einstein equation remains unchanged with respect to the ordinary Einstein equation, the right-hand side changes quite dramatically. Namely, the classical stress-energy tensor T_{ab} is replaced by a the expectation value of a quantum observable $:T_{ab}:$ in a certain state ω. For this expression to be mathematically consistent, we need to require the conservation of the quantum stress-energy tensor, *i.e.*, $\nabla^a:T_{ab}: = 0$.

---

[1]Remember that we chose units such that $8\pi G = c = 1$.

7.2.1   The stress-energy tensor of the Klein–Gordon field

We do not aim at discussing the semiclassical Einstein equation in all possible generality. Instead we restrict our discussion to the semiclassical Einstein equation sourced by a scalar field. The classical *stress-energy tensor of a Klein–Gordon field* φ with equation of motion

$$\mathrm{P}\varphi = (\Box + \xi R + m^2)\varphi = 0$$

may be written as [122]

$$T_{ab} \doteq \tfrac{1}{2}\nabla_a\nabla_b\varphi^2 + \tfrac{1}{4}g_{ab}\Box\varphi^2 - \varphi\nabla_a\nabla_b\varphi + \tfrac{1}{2}g_{ab}g^{cd}\varphi\nabla_c\nabla_d\varphi$$
$$+ \xi(G_{ab} - \nabla_a\nabla_b - g_{ab}\Box)\varphi^2 - \tfrac{1}{2}g_{ab}m^2\varphi^2. \tag{7.2}$$

It may be obtained by varying the classical action of the scalar field with respect to the metric [216, App. E]. The *quantum stress-energy tensor* is obtained from the classical expression (7.2) by replacing products of classical fields by Wick products of quantum fields, *i.e.*,

$$:T_{ab}: \doteq \tfrac{1}{2}\nabla_a\nabla_b:\widehat{\varphi}^2: + \tfrac{1}{4}g_{ab}\Box:\widehat{\varphi}^2: - :\widehat{\varphi}\,\nabla_a\nabla_b\widehat{\varphi}: + \tfrac{1}{2}g_{ab}g^{cd}:\widehat{\varphi}\,\nabla_c\nabla_d\widehat{\varphi}:$$
$$+ \xi(G_{ab} - \nabla_a\nabla_b - g_{ab}\Box):\widehat{\varphi}^2: - \tfrac{1}{2}g_{ab}m^2:\widehat{\varphi}^2:. \tag{7.3}$$

This expression is not obviously conserved as

$$\nabla^a:T_{ab}: = -:(\nabla_b\widehat{\varphi})(\mathrm{P}\widehat{\varphi}):$$

is not necessarily vanishing even if the Wick square was a solution of the equations of motion. Nevertheless, either by a judicious choice in the renormalization freedom of $:\varphi^2:$ and $:\varphi\nabla_a\nabla_b\varphi:$ [122] or, equivalently, by a redefinition of the quantum stress-energy tensor [154], a conserved quantum stress-energy tensor can be found; here we follow the approach of Hollands and Wald. While the renormalization freedom can be used to find a conserved stress-energy tensor, it is not possible to impose the equations of motions on a locally covariant normal ordering prescription [122].

7.2.2   Renormalization of the stress-energy tensor

The *renormalization freedom* of $:\widehat{\varphi}^2:$ and $:\widehat{\varphi}\,\nabla_a\nabla_b\widehat{\varphi}:$ is spanned by m^2, R and

$$g_{ab}m^4,\ g_{ab}m^2R,\ m^2R_{ab},\ \nabla_a\nabla_bR,\ g_{ab}\Box R,\ \Box R_{ab},\ g_{ab}R^2,$$
$$RR_{ab},\ R_{ac}R^c{}_b,\ g_{ab}R_{cd}R^{cd},\ R^{cd}R_{cadb},\ g_{ab}R_{cdef}R^{cdef}.$$

We have to split this renormalization freedom into two classes: (a) combinations of terms that are conserved and represent a true renormalization freedom, and (b) combinations of terms that are not conserved and need to be fixed to produce a conserved $:T_{ab}:$.

Denote by I_{ab} and J_{ab} the two conserved curvature tensors of derivative order 4:

$$I_{ab} \doteq 2RR_{ab} - 2\nabla_a\nabla_bR - \tfrac{1}{2}g_{ab}\left(R^2 + 4\Box R\right),$$
$$J_{ab} \doteq 2R^{cd}R_{cadb} - \nabla_a\nabla_bR - \Box R_{ab} - \tfrac{1}{2}g_{ab}\left(R_{cd}R^{cd} + \Box R\right).$$

The following is often stated in the form of a conjecture (*e.g.*, in [219]):

Proposition 7.1. I_{ab} and J_{ab} span the whole space of conserved fourth order local curvature tensors.

Proof. It is an easy task to confirm this statement by a direct computation along the lines of [68]: Taking the linear span of all fourth order curvature tensors

$$\nabla_a\nabla_b R,\ g_{ab}\Box R,\ \Box R_{ab},\ g_{ab}R^2,\ RR_{ab},\ R_{ac}R^c{}_b,$$
$$g_{ab}R_{cd}R^{cd},\ R^{cd}R_{cadb},\ g_{ab}R_{cdef}R^{cdef},$$

one can show that any covariantly conserved combination C_{ab} must be of the form

$$C_{ab} = \alpha_1\nabla_a\nabla_b R - \alpha_2 g_{ab}\Box R + 2(\alpha_1+\alpha_2)\Box R_{ab} - \tfrac{1}{4}(\alpha_1+2\alpha_2)g_{ab}R^2$$
$$+ (\alpha_1+2\alpha_2)RR_{ab} + (\alpha_1+\alpha_2)g_{ab}R_{cd}R^{cd} - 4(\alpha_1+\alpha_2)R^{cd}R_{cadb},$$

i.e., one obtains (for general metrics) a two-dimensional solution space. For $\alpha_1 = -2, \alpha_2 = 2$ and $\alpha_1 = -1, \alpha_2 = 1/2$ we recover the tensors I_{ab} and J_{ab}, respectively. □

Remark 7.2. *In conformally flat spacetimes (e.g., FLRW spacetimes) the Weyl tensor vanishes and thus the solution space reduces to one dimension as the two tensors become proportional: $I_{ab} = 3J_{ab}$. On the level of traces this proportionality holds for all metrics, namely, $I^a{}_a = 3J^a{}_a = -6\Box R$.*

We therefore find that the *conserved* renormalization freedom of $:T_{ab}:$ is spanned by $m^4 g_{ab}, m^2 G_{ab}, I_{ab}, J_{ab}$. The remaining terms renormalization parameters are fixed by the requirement of $:T_{ab}:$ to be conserved.

7.2.3   Point-splitting regularization of the stress-energy tensor

Up to the renormalization freedom, a normal ordering prescription for the stress-energy tensor is given by the *Hadamard point-splitting method*. Given two linear (possibly tensorial) differential operators $\mathscr{D}_1, \mathscr{D}_2$, the Hadamard point-splitting method yields the expectation value of $:(\mathscr{D}_1\widehat{\varphi})(\mathscr{D}_2\widehat{\varphi}):$ by seperating points, regularizing and then taking the coincidence limit, that is

$$\omega\big(:(\mathscr{D}_1\widehat{\varphi})(\mathscr{D}_2\widehat{\varphi}):\big) = \lim_{x'\to x}\mathscr{D}_1\mathscr{D}_2'\big(\omega_2(x,x') - \mathscr{H}(x,x')\big) = [\mathscr{D}_1\mathscr{D}_2'w],$$

where $\mathscr{D}_2'$ acts on x' and is (implicitly) parallel transported during the limit $x' \to x$.

In the Hadamard point-splitting approach the stress-energy tensor in a state ω of sufficient regularity is thus calculated as

$$\omega(:T_{ab}:) = \frac{1}{8\pi^2}\big(\mathscr{T}_{ab}[w] + \mathscr{T}_{ab}{}^{cd}[\nabla_c\nabla_d w]\big) + \frac{1}{4\pi^2}[v_1]g_{ab}$$
$$+ c_1 m^4 g_{ab} + c_2 m^2 G_{ab} + c_3 I_{ab} + c_4 J_{ab},$$

where $\mathscr{T}_{ab}$ and $\mathscr{T}_{ab}{}^{cd}$ are the differential operators acting, respectively, on $:\widehat{\varphi}^2:$ and $:\widehat{\varphi}\,\nabla_a\nabla_b\widehat{\varphi}:$ in (7.3):

$$\mathscr{T}_{ab} \doteq \tfrac{1}{2}\nabla_a\nabla_b + \tfrac{1}{4}g_{ab}\Box + \xi(G_{ab} - \nabla_a\nabla_b - g_{ab}\Box) - \tfrac{1}{2}g_{ab}m^2,$$
$$\mathscr{T}_{ab}{}^{cd} \doteq -\delta_a^c\delta_b^d + \tfrac{1}{2}g_{ab}g^{cd}.$$

Furthermore, c_i are dimensionless (renormalization) constants fixed once for all spacetimes[2] and the addition of the Hadamard coefficient $[v_1]$ (see (6.6) for an explicit expression) makes the quantum stress-energy tensor conserved because

$$\lim_{x'\to x} \nabla'_a \mathrm{P}\mathscr{H}(x,x') = -\frac{1}{4\pi^2}\nabla_a[v_1].$$

Observe that $c_1 m^4 g_{ab}$ can be interpreted as a renormalization of the cosmological constant and $c_2 m^2 G_{ab}$ corresponds to a renormalization of Newton's gravitational constant G; the remaining two terms have no classical interpretation.

7.2.4   Trace of the stress-energy tensor

Because of its simple form, a first step towards analyzing the stress-energy tensor of a scalar field is often the study its trace, which is given by

$$:T: \doteq g^{ab} :T_{ab}: = -m^2 :\widehat{\varphi}^2: + 3\left(\tfrac{1}{6} - \xi\right)\Box :\widehat{\varphi}^2: - :\widehat{\varphi}\,\mathrm{P}\widehat{\varphi}:. \tag{7.4}$$

It follows that the trace of the stress-energy tensor is calculated via point-splitting as

$$\begin{aligned}
\omega(:T:) &= -\left(m^2 - 3\left(\tfrac{1}{6}-\xi\right)\Box\right)\frac{1}{8\pi^2}[w] + \frac{1}{4\pi^2}[v_1] \\
&\quad + 4c_1 m^4 - c_2 m^2 R - (6c_3 + 2c_4)\Box R,
\end{aligned} \tag{7.5}$$

where c_i are the same constants as above and we used

$$\lim_{x'\to x} \mathrm{P}\mathscr{H}(x,x') = -\frac{3}{4\pi^2}[v_1].$$

Equations (7.4) and (7.5) clearly show the so-called *trace anomaly* [219]. Namely, because the normally ordered quantum field does not satisfy the equations of motion, the massless, conformally coupled scalar field ($m=0$ and $\xi = 1/6$) has non-vanishing trace of the stress-energy tensor although it is conformally invariant. It is not possible to remove the trace anomaly by a judicious choice of the renormalization constants because $[v_1]$ is not a polynomial of m^4, $m^2 R$ and $\Box R$. The trace anomaly is a distinct feature of the quantum theory and does not appear in a classical theory because the classical fields are solutions of the equation of motion.

7.3   The semiclassical Friedmann equations

On FLRW spacetimes (M, g) the classical Einstein equation (2.9) simplifies significantly to the first and second Friedmann equation (2.18) and (2.19). Since the left-hand side remains unchanged when crossing over to the semiclassical Einstein equation, also the semiclassical equations must simplify in an analogue way for every state that satisfies the equation. Whence one obtains the semiclassical Friedmann equations

$$6H^2 + \overline{R} = 2\omega(:\rho:) + 2\Lambda, \tag{7.6a}$$

$$6(\dot{H} + H^2) = -\omega(:\rho: + 3:p:) + 2\Lambda = -\omega(:T: + 2:\rho:) + 2\Lambda, \tag{7.6b}$$

---

[2]c_1, c_2 are due to the renormalization of $:\widehat{\varphi}^2:$ and c_3, c_4 correspond to the renormalization freedom of $:\widehat{\varphi}\,\nabla_a\nabla_b\widehat{\varphi}:$

where the *quantum energy-density* $:\rho:$ and the *quantum pressure* $:p:$ are constructed out of the quantum stress-energy tensor just like their classical analogues are obtained from the classical stress-energy tensor. Henceforth we will restrict again to flat FLRW spacetimes but similar statements can also be made in the case of elliptic and hyperbolic spatial sections.

States that satisfy the semiclassical Einstein equation need to respect the symmetries of the spacetime. Therefore, any candidate state for a solution of the semiclassical Friedmann equations must be homogeneous and isotropic. That is, under reasonable assumptions, it must be a state of the form discussed in Sect. 6.2.2. Important examples of homogeneous and isotropic states are the adiabatic states (Sect. 6.2.3) and the states of low energy (Sect. 6.2.5).

7.3.1   Semiclassical Friedmann equations for the scalar field

For the scalar field, the energy density and pressure are obtained from (7.3) and they read

$$:\rho: = \left((\tfrac{1}{2} - \xi)\partial_t^2 - (\tfrac{1}{4} - \xi)\Box + 3\xi H^2 + \tfrac{1}{2}m^2 \right):\widehat{\varphi}^2: - :\widehat{\varphi}\,(\partial_t^2 - \tfrac{1}{2}\Box)\widehat{\varphi}:, \qquad (7.7)$$

$$3:p: = \left((\tfrac{1}{2} - \xi)\partial_t^2 + (\tfrac{1}{4} - 2\xi)\Box - \xi(6\dot{H} + 9H^2) - \tfrac{3}{2}m^2 \right):\widehat{\varphi}^2: - :\widehat{\varphi}\,(\partial_t^2 + \tfrac{1}{2}\Box)\widehat{\varphi}:$$

with respect to cosmological time t. The expressions for conformal time τ are given by the replacement $\partial_t \mapsto a^{-1}\partial_\tau$. A short calculation shows that the difference $3:p: - :\rho:$ agrees with (7.4).

The expectation values of $:\rho:$ and $:p:$ in a state ω can again be calculated via Hadamard point-splitting. For the energy density this approach yields:

$$\begin{aligned}
\omega(:\rho:) = \mathscr{P}[w] - \lim_{x' \to x} \left[(\partial_t^2 - \tfrac{1}{2}\Box)w \right] - \frac{1}{4\pi^2}[v_1] \\
- c_1 m^4 + 3c_2 m^2 H^2 - 6(3c_3 + c_4)(2H\ddot{H} - \dot{H}^2 + 6H^2\dot{H}),
\end{aligned} \qquad (7.8)$$

where we have used the differential operator

$$\mathscr{P} \doteq (\tfrac{1}{2} - \xi)\partial_t^2 - (\tfrac{1}{4} - \xi)\Box + 3\xi H^2 + \tfrac{1}{2}m^2$$

and the renormalization constants c_i are again the same as in Sects. 7.2.3 and 7.2.4. The coincidence limit of the Hadamard coefficient v_1 on FLRW spacetimes can be obtain from (6.6):

$$\begin{aligned}
2[v_1] = \tfrac{1}{4}m^4 - 3(\tfrac{1}{6} - \xi)(\dot{H} + 2H^2)m^2 + 9(\tfrac{1}{6} - \xi)^2(\dot{H}^2 + 4\dot{H}H^2 + 4H^4) \\
- \tfrac{1}{30}(\dot{H}H^2 + H^4) + \tfrac{1}{12}(\tfrac{1}{5} - \xi)\Box R.
\end{aligned} \qquad (7.9)$$

The point-split expression for $\omega(:p:)$ will not concern us here and is left as an exercise to the reader.

The next step is to replace the Hadamard point-splitting prescription with something a that is slightly more useful under the given circumstances.

7.3.2   Adiabatic regularization

An effective means of regularizing in momentum space is provided by the *adiabatic regularization* [50, 173], which is essentially equivalent to the Hadamard point-splitting regularization discussed above. In the adiabatic regularization prescription one subtracts from a homogeneous and isotropic state ω the bidistribution specified by the adiabatic modes $W_k^{(n)}$ of sufficiently high order n. The usefulness of this prescription lies in the fact that the bidistributions constructed out of the adiabatic modes satisfy the microlocal spectrum condition in the Sobolev sense up to a certain order. Consequently, the two regularization prescriptions can only differ by local curvature tensors because both the Hadamard parametrix and the adiabatic modes are constructed from the local geometric structure of the spacetime.

It is therefore necessary to find the difference between point-splitting and adiabatic regularization

$$\lim_{x' \to x} \mathscr{D} \left(\mathscr{H}_n(x, x') - \frac{1}{(2\pi)^3 a(\tau)a(\tau')} \int_{\mathbb{R}^3} \overline{W}_k^{(m)}(\tau) W_k^{(m)}(\tau') \, \mathrm{e}^{\mathrm{i}\vec{k}\cdot(\vec{x}-\vec{x}')} \, \mathrm{d}\vec{k} \right) \qquad (7.10)$$

for various (bi)differential operators $\mathscr{D}$ on $C^\infty(M \times M)$ and the minimal orders n, m depending on the order of $\mathscr{D}$. Note that we omitted the necessary ε-regularization in the integrand.

It is helpful to note that the (truncated) Hadamard parametrix on FLRW is spatially isotropic and homogeneous and therefore $\mathscr{H}(\tau, \vec{x}; \tau', \vec{x}') = \mathscr{H}(\tau, \tau', |\vec{x} - \vec{x}'|)$. This fact can simplify some computations because the coincidence limit can be taken in two steps: first one takes the limit onto the equal time surface $\tau = \tau'$ and then the spatial coincidence limit $\vec{x}' \to \vec{x}$. Making efficient use of the equation of motion, this has the advantage that we can replace any higher than first order time derivative in $\mathscr{D}$ by a spatial derivative and one needs to calculate the temporal coincidence limit only with the differential operators ∂_τ, ∂_τ' and $\partial_\tau \partial_\tau'$. A proof of this statement can be found in [196, Chap. 5].

The computation (7.10) can be done in a general and efficent manner with a computer algebra system by combining the method of Avramidi to calculate Hadamard coefficients (Sect. 1.4.4), the coordinate expansion of the world function (Sect. 1.4.5) and analytic Fourier transformation. See also [83] for a related approach or [196] for a different method that makes more efficient use of the symmetries of FLRW spacetimes.

The difference between point-splitting and adiabatic regularization for the Wick square can be calculated in this way as

$$\lim_{x' \to x} \left(\mathscr{H}_0(x, x') - \frac{1}{(2\pi)^3 a^2} \int_{\mathbb{R}^3} \frac{1}{2\omega_k} \mathrm{e}^{\mathrm{i}\vec{k}\cdot(\vec{x}-\vec{x}')} \, \mathrm{d}\vec{k} \right)$$
$$= \frac{B}{16\pi^2} \left(1 - 2\gamma - \ln\left(\tfrac{1}{2}\lambda^2 B\right) \right) + \frac{R}{288\pi^2}, \qquad (7.11)$$

where we used the potential $B = m^2 + (\xi - \tfrac{1}{6})R$ and γ denotes Euler's constant.

The exact form of the mode subtraction performed in the adiabatic regularization is inessential as long as it has the right $\vec{k}$-asymptotics (*cf.* [196, Chap. 5]). For example, instead of subtracting the adiabatic modes of order zero in (7.11), one can perform

the following subtraction (*cf.* [178, 3] and Sect. 8.1.4):

$$\lim_{x'\to x}\left(\mathscr{H}_0(x,x') - \frac{1}{(2\pi)^3 a^2}\int_{\mathbb{R}^3}\left(\frac{1}{2\omega_k(\eta)} - \frac{\omega_k^2 - \omega_k(\eta)^2}{4\omega_k(\eta)^3}\right)e^{i\vec{k}\cdot(\vec{x}-\vec{x}')}\,d\vec{k}\right)$$

$$= \frac{B(\eta)a(\eta)^2}{16\pi^2 a^2} - \frac{B}{16\pi^2}\left(2\gamma + \ln\frac{\lambda^2 a(\eta)^2 B(\eta)}{2a^2}\right) + \frac{R}{288\pi^2}, \qquad (7.12)$$

where η is an arbitrary instant in conformal time. But this realization is much more important once one attempts to calculate the adiabatic subtraction with several derivatives such as the energy density, where higher order adiabatic modes need to be subtracted. Here one generically finds that the adiabatic regularization involves terms with higher derivatives of the metric that do not become singular in the coincidence limit, *i.e.*, terms which do not need to be subtracted to achieve the regularization.

To find an adiabatic subtraction for other components of the stress-energy tensor one can follow the approach of Bunch [50] and take the second order adiabatic modes but discard all (non-singular) terms involving higher than fourth order time-derivatives of the metric. Comparing this subtraction scheme with the Hadamard point-splitting regularization for the energy density, we find (see also [114])

$$\lim_{x'\to x}\left(\mathscr{D}_\rho\, a(\tau)a(\tau')\mathscr{H}_1(x,x') - \frac{1}{(2\pi)^3 a^2}\int_{\mathbb{R}^3}e^{i\vec{k}\cdot(\vec{x}-\vec{x}')}\left(\frac{k}{2a^2} + \frac{m^2 - 6(\xi - \frac{1}{6})H^2}{4k}\right.\right.$$

$$\left.\left. - \frac{m^4 a^2 + 12(\xi - \frac{1}{6})m^2 a^2 H^2 + 36(\xi - \frac{1}{6})^2(6H^2\dot{H} - \dot{H}^2 + 2H\ddot{H})}{16k^3}\right)d\vec{k}\right)$$

$$= -\frac{1}{4\pi^2}[v_1] - \frac{H^4}{960\pi^2} + \left(2 - 2\gamma + \ln\frac{a^2}{2\lambda^2}\right)\frac{m^4}{64\pi^2}$$

$$+ \left(1 + 18(\xi - \tfrac{1}{6})\left(2 - 2\gamma + \ln\frac{a^2}{2\lambda^2}\right)\right)\frac{m^2 H^2}{96\pi^2} + \frac{3(\xi - \frac{1}{6})^2 H^2 R}{8\pi^2}$$

$$+ \left(\frac{1}{17280\pi^2} - \frac{\xi - \frac{1}{6}}{288\pi^2} - \frac{(\xi - \frac{1}{6})^2}{32\pi^2}\left(2 - 2\gamma + \ln\frac{a^2}{2\lambda^2}\right)\right)I_0{}^0, \qquad (7.13)$$

where the (bi)differential operator

$$2a^4\mathscr{D}_\rho = a^2\big(m^2 + (1 - 6\xi)H^2\big)\boxtimes\mathrm{id} + 2(6\xi - 1)aH\partial_\tau\boxtimes\mathrm{id} + \partial_\tau\boxtimes\partial_\tau + \sum_{i=1}^{3}\partial_{x^i}\boxtimes\partial_{x^i}$$

can be derived from (7.7). In the case of conformal coupling it is in fact sufficient to work with adiabatic modes of order zero for this computation and many of the terms in the above formula drop out.

Observe that both the point-splitting regularization with the truncated Hadamard parametrix and the adiabatic regularization do not depend on arbitrarily high derivatives in the metric. Consequently, it is possible to perform both regularization schemes in non-smooth spacetimes. This will be essential in Chap. 8.

8

Solutions of the semiclassical Einstein eq.

Introduction

If one wants to attach any physical meaning to the semiclassical Einstein equation (7.1), it is necessary that solutions of this equation exist and that it possesses a well-posed initial value problem. It is not difficult to show that solutions do indeed exist at least in two very special scenarios: Minkowski spacetime and de Sitter spacetime. In both cases the semiclassical Einstein equation (or, alternatively, the semiclassical Friedmann equations) can be solved for a specific choice of the renormalization constants.

Solutions of the semiclassical Friedmann equations were investigated already numerically by Anderson in a series of four articles beginning with the massless conformally coupled scalar field [12, 13] and later also considering the massive field [14, 15]. Anderson discovered a complex landscape of solutions depending on the choice of renormalization constants and studied in particular solutions which show an asymptotically classical behaviour at late times.

More recently, Pinamonti discussed the local existence of solutions to the semiclassical Friedmann equations in so-called null Big Bang (NBB) spacetimes [178], where initial values are specified on the initial lightlike singularity.

In [3] the author and Pinamonti proved for the first time the existence of global solutions to the semiclassical Einstein equation coupled to a massive, conformally coupled scalar field in 'non-trivial' spacetimes. More precisely, it was shown that the semiclassical Friedmann equations can solved simultaneously for the spacetime metric (*i.e.*, the scale factor or the Hubble function) and a quantum state from initial values given at some Cauchy surface. This was achieved by showing existence and uniqueness of local solutions for given initial values and subsequently extending local solutions to a maximal solution that cannot be extended any further because it exists either eternally or reaches a singularity. In this chapter, a slightly updated version of the results of [3] will be presented and complemented with recent numerical results.

In more generality, solving the semiclassical Einstein equations for a given quantum field means the following:

Given initial values for a spacetime metric and a quantum state prescribed on a three-dimensional Riemannian manifold (Σ, h)*, do there exist a globally hyperbolic manifold* (M, g) *of which* (Σ, h) *is a Cauchy surface and a state* ω *(preferably a Hadamard state) such that the semiclassical Einstein equation* (7.1) *is fulfilled?*

In a concrete case this problem can be tackled by selecting a class of globally hyperbolic spacetimes that are foliated by the same topological Cauchy surface and a functional that associates to each spacetime in this class a unique states. For this approach to succeed, it is expected that the mentioned functional must satisfy some minimal regularity conditions with respect to the metric, *e.g.*, continuous differentiability.

8.1   Preliminaries

8.1.1   The traced stress-energy tensor

Recall from Sect. 7.2.4 that the expectation value of the traced stress-energy tensor for a massive, conformally coupled scalar field reads

$$\omega(:T:) = -m^2\omega(:\widehat{\varphi}^2:) + \frac{1}{4\pi^2}[v_1] - (6c_3 + 2c_4)\Box R \tag{8.1}$$

$$= -\frac{m^2}{8\pi^2}[w] + \frac{1}{4\pi^2}[v_1] + 4c_1 m^4 - c_2 m^2 R - (6c_3 + 2c_4)\Box R,$$

where the Hadamard coefficient $[v_1]$ is obtained from (7.9) with $\xi = 1/6$ as

$$[v_1] = -\frac{H^2}{60}(\dot{H} + H^2) - \frac{1}{720}\Box R + \frac{m^4}{8}.$$

Working with the traced stress-energy tensor $:T:$ simplifies calculations considerably compared to the energy density $:\rho:$ given by (7.8).

In order to find solutions of the semiclassical Friedmann equation with the methods discussed here, it is necessary to fix the renormalization constants according to the following rules:

We will choose c_3, c_4 in such a way as to cancel higher order derivatives of the metric coming from $[v_1]$. Following [218] and [217, Chap. 4.6], this is necessary because we want to have a well-posed initial value problem for a second-order differential equation. Removing the $\Box R$ term might not be suitable for describing the physics close to the initial Big Bang singularity. In the Starobinsky model this term is the single term which is considered to drive a phase of rapid expansion close to the Big Bang, see the original paper of Starobinsky [201], its further development [141] and also [112, 114] for a recent analysis. However, this is surely suitable to describe the physics in the regime where $\Box H \ll H^4$.

Furthermore, remember that changing c_1 corresponds to a renormalization of the cosmological constant Λ, whereas a change of c_2 corresponds to a renormalization of the Newton constant G (cf. Sect. 7.2.3). For this reason we choose c_1 in such a way that no contribution proportional to m^4 is present in $\omega(:T:)$ and we set c_2 in order to cancel the terms proportional to $m^2 R$ in $\omega(:T:)$. All in all, we fix the renormalization constants as

$$4c_1 = -\frac{1}{32\pi^2}, \quad c_2 = \frac{1}{288\pi^2} \quad \text{and} \quad 6c_3 + 2c_4 = -\frac{1}{2880\pi^2}.$$

8.1.2   The semiclassical Friedmann equations

We can rewrite the semiclassical Friedmann equations to make use of the simplicity of the traced stress-energy tensor for the conformally coupled scalar field: Adding the equations (7.6) (for flat FLRW spacetimes) yields

$$-6(\dot{H} + 2H^2) = \omega(:T:) - 4\Lambda. \tag{8.2}$$

Since $:T: = 3:p: - :\rho:$, this equation is equivalent to (7.6) if we also specify an initial value $\rho_0 \doteq \omega(:\rho:)(\tau_0)$ for the expectation value of the energy density at a time τ_0:

$$3H_0^2 \doteq 3H(\tau_0)^2 = \rho_0 + \Lambda. \tag{8.3}$$

We call (8.3) the *constraint equation* because it relates the initial value H_0 for the spacetime geometry with the initial value ρ_0 for its matter content; these values cannot be fixed independently.

Inserting (8.1) into (8.2) and solving for $\dot{H}$ we thus find

$$\dot{H} = \frac{1}{H_c^2 - H^2}\left(H^4 - 2H_c^2 H^2 - \tfrac{15}{2}m^4 + 240\pi^2(m^2\omega(:\widehat{\varphi}^2:) + 4\Lambda)\right), \tag{8.4}$$

which can integrated in conformal time to give

$$H(\tau) = H_0 + \int_{\tau_0}^{\tau} \frac{a(\eta)}{H_c^2 - H(\eta)^2}\left(H(\eta)^4 - 2H_c^2 H(\eta)^2 - \tfrac{15}{2}m^4\right. \tag{8.5}$$
$$\left. + 240\pi^2(m^2\omega(:\widehat{\varphi}^2:)(\eta) + 4\Lambda)\right)\mathrm{d}\eta,$$

where $H_c^2 \doteq 1440\pi^2/(8\pi G) = 180\pi/G$. This integral equation will be our principal tool to solve the semiclassical Einstein equation.

8.1.3 A choice of states

As discussed in the introduction, a possible approach to solving the semiclassical Einstein equation is to select a class of candidate spacetimes and then for each of these spacetimes a unique state. Here we restrict ourselves to the semiclassical Friedmann equations as given by (8.2) and (8.3), *viz.*, the candidate spacetimes are flat FLRW spacetimes. It remains to find a functional that associates to each flat FLRW spacetime a suitable state.

It would be desirable to associate to each spacetime a Hadamard state. In the literature there are a few concrete examples of such states but unfortunately none of them are suitable for our purposes. On FLRW spacetimes there is the notable construction of states of low energy discussed in Sect. 6.2.5, which are also Hadamard. But the employed construction is based on a smearing of the modes with respect to an extended function of time and *a priori* we do not even know if a solution of (8.5) exists in the future of the initial Cauchy surface. The holographic constructions of Hadamard states, discussed in Sect. 6.3, requires that the spacetime has certain asymptotic properties which are not under control for generic FLRW spacetimes.

Moreover, for technical reasons to be discussed later, we also have to consider spacetimes with C^1 metrics. But on spacetimes with non-smooth metrics Hadamard states cannot exist. Instead we will use the construction of adiabatic states of order zero as presented in Sect. 6.2.4, which is also applicable to spacetimes with low regularity. The price we have to pay for working with non-Hadamard states is that the solutions of (8.5) are not be smooth spacetimes.

We recall, that the states constructed in Sect. 6.2.4 are of the form

$$\omega_2(x, x') = \frac{1}{(2\pi)^3 a(\tau)a(\tau')}\int_{\mathbb{R}^3} \overline{\chi}_k(\tau)\chi_k(\tau')\, \mathrm{e}^{\mathrm{i}\vec{k}\cdot(\vec{x}-\vec{x}')}\,\mathrm{d}\vec{k} \tag{8.6}$$

with modes $\chi_k = \sum_n \chi_k^{(n)}(\tau)$ given by

$$\chi_k^{(0)}(\tau) = \frac{1}{\sqrt{2k_0}} e^{ik_0\tau}, \quad \chi_k^{(n)}(\tau) = \int_{\tau_0}^{\tau} \frac{\sin\left(k_0(\eta - \tau)\right)}{k_0} V(\eta) \chi_k^{(n-1)}(\eta) \, d\eta,$$

where $k_0^2 = k^2 + a(\tau_0)^2 m^2$ and $V(\tau) = m^2(a(\tau)^2 - a(\tau_0)^2)$. Note that χ_k and ω_2 can be defined in this way even if the scale factor a is only C^1. This may be confirmed by taking a closer look at Prop. 6.2.

8.1.4   Adiabatic regularization of the Wick square

The integral equation (8.5) does not contain the two-point function but instead only its smooth part w in the coincidence limit, *i.e.*, Hadamard point-splitting has to applied to (8.6).

The equation (8.5) that we seek to solve contains the Wick square $:\widehat{\varphi}^2:$ in a state ω and thus (on a smooth spacetime) we would need to compute

$$\omega(:\widehat{\varphi}^2:) = \lim_{x' \to x} \left(\omega_2(x, x') - \mathscr{H}(x, x')\right) - 4c_1 m^4 + c_2 m^2 R. \tag{8.7}$$

Since we are on a FLRW spacetime, we can use the method of adiabatic regularization instead (Sect. 7.3.2) to perform an equivalent subtraction on the level of modes. The difference of the two approaches is given in (7.11) or, equivalently, (7.12). It is useful to show directly that this regularization prescription indeed regularizes the two-point distribution (8.6):

Proposition 8.1. *The regularized two-point distribution*

$$\omega_2(\tau, \vec{x} - \vec{x}') - \lim_{\varepsilon \to 0^+} \frac{1}{(2\pi)^3 a(\tau)^2} \int_{\mathbb{R}^3} \left(\frac{1}{2k_0} - \frac{V(\tau)}{4k_0^3}\right) e^{i\vec{k}\cdot(\vec{x}-\vec{x}')} e^{-\varepsilon k} \, d\vec{k},$$

with ω_2 given by (8.6), converges in the coinciding point limit for all continuously differentiable scale factors a.

Proof. We have to show that

$$\lim_{\varepsilon \to 0^+} \int_{\mathbb{R}^3} \left(|\chi_k|^2 - \frac{1}{2k_0} + \frac{V}{4k_0^3}\right) e^{-\varepsilon k} \, d\vec{k} = \int_{\mathbb{R}^3} \left(|\chi_k|^2 - \frac{1}{2k_0} + \frac{V}{4k_0^3}\right) d\vec{k} \tag{8.8}$$

is finite. To this end we expand the product $|\chi_k|^2$ with $\chi_k = \sum_n \chi_k^n$ as

$$|\chi_k|^2 = \sum_{n=0}^{\infty} \sum_{l=0}^{n} \chi_k^l \, \overline{\chi}_k^{n-l}$$

in terms of the order n. Inserting this expansion into (8.8), we can prove the statement order by order:

0th order.   Since $\chi_k^0 \overline{\chi}_k^0 = (2k_0)^{-1}$, the first term in the subtraction exactly cancels the zeroth order term $|\chi_k^0|^2$ in (8.8).

1st order.   Using an integration by parts, we can rewrite the first order terms as

$$(\chi_k^0 \overline{\chi}_k^1 + \chi_k^1 \overline{\chi}_k^0)(\tau) = \frac{1}{k_0^2} \int_{\tau_0}^{\tau} \sin\left(k_0(\eta - \tau)\right) \cos\left(k_0(\eta - \tau)\right) V(\eta) \, d\eta$$

$$= \frac{1}{2k_0^2} \int_{\tau_0}^{\tau} \sin\left(2k_0(\eta - \tau)\right) V(\eta) \, d\eta$$

$$= -\frac{V(\tau)}{4k_0^3} + \frac{1}{4k_0^3} \int_{\tau_0}^{\tau} \cos\left(2k_0(\eta - \tau)\right) V'(\eta) \, d\eta.$$

While the first summand $V(\tau)(4k_0^3)^{-1}$ in the last line is exactly cancelled by the second term in the subtraction in (8.8), the second summand yields

$$\int_{\mathbb{R}^3} \frac{1}{4k_0^3} \left(\int_{\tau_0}^{\tau} \cos\left(2k_0(\eta - \tau)\right) V'(\eta) \, d\eta \right) e^{-\varepsilon k} \, d\vec{k}$$

$$= \pi \int_0^{\infty} \frac{k^2}{k_0^3} \left(\int_{\tau_0}^{\tau} \cos\left(2k_0(\eta - \tau)\right) V'(\eta) \, d\eta \right) e^{-\varepsilon k} \, dk$$

$$= \pi \int_{a_0 m}^{\infty} k_0^{-1} \sqrt{1 - a^2 k_0^{-2}} \left(\int_{\tau_0}^{\tau} \cos\left(2k_0(\eta - \tau)\right) V'(\eta) \, d\eta \right) e^{-\varepsilon k} \, dk_0$$

$$= \pi \int_{\tau_0}^{\tau} V'(\eta) \left(\int_{a_0 m}^{\infty} k_0^{-1} \cos\left(2k_0(\eta - \tau)\right) e^{-\varepsilon k} \, dk_0 \right) d\eta - R(\tau) \qquad (8.9)$$

for $\varepsilon > 0$. Here R is a finite remainder term since it contains terms in the k_0-integration which decay at least like k_0^{-3}. Notice that, in the last equation of the previous formula, thanks to the positivity of ε we have switched the order in which the k_0- and η-integration are taken. We would like to show that the weak limit $\varepsilon \to 0^+$ can be taken before the η-integration in (8.9).

To this end it remains to be shown that the k_0-integral in (8.9) converges in the limit $\varepsilon \to 0^+$ to an integrable function in $[\tau_0, \tau]$. First, note that the exponential integral

$$E_1(x) = \Gamma(0, x) = \int_1^{\infty} \frac{e^{-xt}}{t} \, dt = \int_0^1 \frac{e^{-x}}{x - \ln(1 - s)} \, ds \qquad (8.10)$$

converges for $x \neq 0, \operatorname{Re} x \geq 0$. To show the identity, we used the substitution

$$t = -x^{-1} \ln(1 - s) + 1$$

involving a subtle but inconsequential change of the integration contour in the complex plane if x is complex. Then we easily see that

$$\lim_{\varepsilon \to 0^+} \int_{a_0 m}^{\infty} k_0^{-1} e^{\pm 2ik_0(\eta - \tau) - \varepsilon k_0} \, dk_0 = E_1\left(\pm 2ia_0 m(\eta - \tau)\right) \qquad (8.11)$$

converges for $\eta \neq \tau$. This result is related with (8.9) via

$$\lim_{\varepsilon \to 0^+} \int_{a_0 m}^{\infty} k_0^{-1} \cos\left(2k_0(\eta - \tau)\right) e^{-\varepsilon k} \, dk_0 = \lim_{\varepsilon \to 0^+} \int_{a_0 m}^{\infty} k_0^{-1} \cos\left(2k_0(\eta - \tau)\right) e^{-\varepsilon k_0} \, dk_0,$$

where we have used the boundedness of $k - (k^2 - a_0^2 m^2)^{1/2}$ and (8.11). Finally, a bound sufficient to see the η-integrability of the k_0-integral in (8.9) can be obtained from the identity in (8.10), namely,

$$\left| E_1(ix) \right| = \left| \int_0^1 (ix - \ln(1-s))^{-1} \, ds \right| \leq \int_0^1 (x^2 + s^2)^{-1/2} \, ds = \ln \left(\frac{1 + \sqrt{1+x^2}}{x} \right).$$

2nd order.   For the second order we calculate

$$
\begin{aligned}
(\chi_k^0 \, \overline{\chi}_k^2 & + \chi_k^1 \, \overline{\chi}_k^1 + \chi_k^2 \, \overline{\chi}_k^0)(\tau) \\
&= \frac{1}{k_0^3} \int_{\tau_0}^{\tau} \sin\left(k_0(\eta - \tau)\right) V(\eta) \\
&\qquad \times \left(\int_{\tau_0}^{\eta} \sin\left(k_0(\xi - \eta)\right) V(\xi) \cos\left(k_0(\xi - \tau)\right) d\xi \right. \\
&\qquad\qquad \left. + \frac{1}{2} \int_{\tau_0}^{\tau} \sin\left(k_0(\xi - \tau)\right) V(\xi) \cos\left(k_0(\xi - \eta)\right) d\xi \right) d\eta \\
&= \frac{1}{k_0^3} \int_{\tau_0}^{\tau} \sin\left(k_0(\eta - \tau)\right) V(\eta) \left(\int_{\tau_0}^{\eta} \sin\left(k_0(2\xi - \eta - \tau)\right) V(\xi) d\xi \right) d\eta \\
&= \frac{1}{2k_0^4} \int_{\tau_0}^{\tau} \sin\left(k_0(\eta - \tau)\right) V(\eta) \left(\int_{\tau_0}^{\eta} \cos\left(k_0(2\xi - \eta - \tau)\right) V'(\xi) d\xi \right. \\
&\qquad\qquad \left. - \cos\left(k_0(\eta - \tau)\right) V(\eta) \right) d\eta,
\end{aligned}
$$
(8.12)

where we have used integration by parts in the last equality. It is easy to obtain a $\vec{k}$-uniform estimate for the integral above and thus the integrability of the second order follows from $\int_{\mathbb{R}^3} k_0^{-4} \, d\vec{k} < \infty$.

Higher orders.   For orders $n > 2$ it is sufficient to use the rough estimate from (6.20):

$$
\begin{aligned}
\left| \sum_{n=3}^{\infty} \sum_{l=0}^{n} \chi_k^l \, \overline{\chi}_k^{n-l} \right|(\tau) &\leq \frac{1}{2k_0} \sum_{n=3}^{\infty} \frac{2^n}{n!} \left(\frac{1}{k_0} \int_{\tau_0}^{\tau} \left| V(\eta) \right| d\eta \right)^3 \left(\int_{\tau_0}^{\tau} (\tau - \eta) \left| V(\eta) \right| d\eta \right)^{n-3} \\
&\leq \frac{4}{k_0^4} \left(\int_{\tau_0}^{\tau} \left| V(\eta) \right| d\eta \right)^3 \exp\left(2 \int_{\tau_0}^{\tau} (\tau - \eta) \left| V(\eta) \right| d\eta \right). \quad (8.13)
\end{aligned}
$$

As above, the integrability of the higher orders follows from $\int_{\mathbb{R}^3} k_0^{-4} \, d\vec{k} < \infty$.

Note that none of the estimates above depends on higher derivatives of the scale factor. Therefore, combining these partial results, we see that the thesis holds true.   □

It follows that we can consistently define the renormalized Wick square of the state given by (8.6) at conformal time τ for every FLRW spacetime with C^1 scale

factor

$$
\omega(:\widehat{\varphi}^2:) = \frac{1}{(2\pi)^3 a^2} \int_{\mathbb{R}^3} \left(|\chi_k|^2 - \frac{1}{2k_0} + \frac{V}{4k_0^3} \right) \mathrm{d}\vec{k}
$$
$$
+ \frac{m^2}{(4\pi)^2} \left(\frac{1}{2} - \left(\frac{a_0}{a(\tau)} \right)^2 + 2\ln\left(\frac{a_0}{a(\tau)} \right) + 2\ln\left(\frac{e^\gamma m\lambda}{\sqrt{2}} \right) \right), \tag{8.14}
$$

which coincides with (8.7) for smooth spacetimes. Moreover, we notice that, as a consequence of the previous proposition, it is possible to obtain global estimates for the renormalized Wick square:

Proposition 8.2. *The renormalized Wick square is bounded on every $a' \in C[\tau_0, \tau_1]$ with $a > 0$ in $[\tau_0, \tau_1]$ for every τ_1 and with $a(\tau_0) = a_0$, namely,*

$$
\left| \omega(:\widehat{\varphi}^2:)(\tau_1) \right| \leq C \left(\sup_{[\tau_0,\tau_1]} a,\ \sup_{[\tau_0,\tau_1]} a',\ (\tau_1 - \tau_0),\ \frac{1}{\inf_{[\tau_0,\tau_1]} a} \right)
$$

where C is a finite increasing function.

Proof. The proof of this proposition and the explicit value of C, can be obtained combining (8.7) with (7.12) and then analyzing the adiabatic subtraction (8.8) order by order as in the proof of the preceding proposition. $\qquad \square$

8.1.5 Adiabatic regularization of the energy density

There is another nice feature about the states we have constructed above. Thanks to the conformal coupling of the scalar field with the curvature, the energy density computed in these states is finite even though these states are (on smooth spacetimes) only adiabatic states of order zero. This is a crucial feature which permits us to solve the constraint (8.3) as a first step towards solving the semiclassical Einstein equation.

Proposition 8.3. *The energy density ρ in the state ω defined by (8.6) at the initial time $\tau = \tau_0$ is finite.*

Proof. Following [114], in order to show that $\rho(\tau_0)$ is finite, we just need to show that the adiabatically regularized expression (see also (7.13) and the subsequent remark)

$$
\int_0^\infty \left(\left(|\chi_k'|^2 + (k^2 + m^2 a^2)|\chi_k|^2 \right) - \left(\left| W_k^{(0)\prime} \right|^2 + (k^2 + m^2 a^2)\left| W_k^{(0)} \right|^2 \right) \right) k^2 \, \mathrm{d}k \tag{8.15}
$$

does not diverge at $\tau = \tau_0$. Evaluating the expression (8.15) at $\tau = \tau_0$ gives

$$
\frac{m^4}{8} \int_0^\infty \frac{a_0^2 (a')^2}{(k^2 + m^2 a_0^2)^{5/2}} k^2 \, \mathrm{d}k < \infty. \qquad \square
$$

Notice that the previous proposition only guarantees that the energy density is well-defined at the initial time. Nevertheless, the conservation equation for the stress-energy tensor permits to state that it is well-defined everywhere.

The expression (8.15) coincides with the energy density ρ of the system up to a conformal rescaling and up to the addition of some finite terms. Thus, since the energy density ρ is finite in the considered state, the constraint (8.3) holds, provided a suitable choice of $H(\tau_0)$ and Λ is made. We stress that, if we do not want to alter Λ, the same result can be achieved adding classical radiation to the energy density of the universe in a suitable state.

We would like to conclude this section with a remark. In adiabatic states of order zero the expectation values of local fields containing derivatives are usually ill-defined. Despite this, in the case of conformal coupling and for our choice of initial conditions (6.18), the energy density turns out to be well-defined. This is essentially due to the fact that in the massless conformally coupled case the adiabatic modes of order zero are solutions of the mode equation (6.10) and in that case the obtained state is the well known conformal vacuum. Hence, the adiabatically regularized energy density vanishes. In the massive case the states constructed above are not very different than the conformal vacuum and, in particular, the energy density remains finite under that perturbation.

8.2   Local solutions

Our aim is to show the existence and uniqueness of local solutions to the semiclassical Friedmann equation. In particular, according to the discussion in the introduction, we will analyze the uniqueness and existence of solutions of (8.5). Similar to the Picard–Lindelöf theorem, we will use the Banach fixed-point theorem to achieve this goal. Some results on functional derivatives and the Banach fixed-point theorem are collected in Chap. 3, in particular Sect. 3.4.

Solving (8.5) is equivalent to finding fixed-points of the functional F defined by

$$
F(H)(\tau) \doteq H_0 + \int_{\tau_0}^{\tau} \frac{a(\eta)}{H_c^2 - H(\eta)^2} \left(H(\eta)^4 - 2H_c^2 H(\eta)^2 - \tfrac{15}{2} m^4 \right.
\tag{8.16a}
$$
$$
\left. + 240\pi^2 (m^2 \omega(:\widehat{\varphi}^2:)(H)(\eta) + 4\Lambda) \right) \mathrm{d}\eta,
$$
$$
\doteq H_0 + \int_{\tau_0}^{\tau} f(H)(\eta)\,\mathrm{d}\eta.
\tag{8.16b}
$$

Since $\omega(:\widehat{\varphi}^2:)(H)$ is well-defined for continuous Hubble functions (see also (8.14)), we select for the Banach space of candidate Hubble functions H the space[1] $C[\tau_0, \tau_1]$, $\tau_0 < \tau_1$, equipped with the uniform norm

$$
\|X\|_{C[\tau_0,\tau_1]} \doteq \|X\|_\infty \doteq \sup_{\tau \in [\tau_0,\tau_1]} |X(\tau)|.
$$

However, once τ_0 and the initial condition $a_0 = a(H)(\tau_0) > 0$ are fixed, we find that

$$
a(H)(\tau) = a_0 \left(1 - a_0 \int_{\tau_0}^{\tau} H(\eta)\,\mathrm{d}\eta \right)^{-1},
\tag{8.17}
$$

---

[1]Until fixed, we take both τ_0 and τ_1 as variable and thus consider a family of Banach spaces.

as a functional of H, is not continuous on $C[\tau_0, \tau_1]$. But we can find an open subset

$$\mathcal{U}[\tau_0, \tau_1] \doteq \{H \in C[\tau_0, \tau_1] \mid \|H\|_{C[\tau_0, \tau_1]} < \min\{a(\tau_0)^{-1}(\tau_1 - \tau_0)^{-1}, H_c\}\} \quad (8.18)$$

on which a and thus also $V = m^2(a^2 - a_0^2)$ depend smoothly on H. Indeed, we can show the following:

Proposition 8.4. *The functional*

$$f(H) = \frac{a(H)}{H_c^2 - H^2}\left(H^4 - 2H_c^2 H^2 - \tfrac{15}{2}m^4 + 240\pi^2(m^2\omega(:\widehat{\varphi}^2:)(H) + 4\Lambda)\right) \quad (8.19)$$

is continuously differentiable on $\mathcal{U}[\tau_0, \tau_1]$ for arbitrary but fixed τ_0, τ_1 and $a_0 = a(\tau_0)$.

Proof. Given (8.7), (7.12), Prop. 8.1 and Prop. 8.2, it is enough to show that $a(H)$ and $(H_c^2 - H^2)^{-1}$ are bounded and that $\omega(:\widehat{\varphi}^2:)(H)(\tau_0)$ is continuously differentiable. The former is assured by the condition $\|H\|_{C[\tau_0, \tau_1]} < \min\{a_0^{-1}(\tau_1 - \tau_0)^{-1}, H_c\}$ in the definition of $\mathcal{U}[\tau_0, \tau_1]$. For the latter it remains to be shown that the renormalized Wick square (8.14) is continuously differentiable on $\mathcal{U}[\tau_0, \tau_1]$:

We start by calculating the functional derivative of the scale factor

$$\mathrm{d}a(H; \delta H)(\tau) = a(H)(\tau)^2 \int_{\tau_0}^{\tau} \delta H(\eta)\,\mathrm{d}\eta.$$

The functional derivatives for a^{-2}, $\ln a$, V and V' follow easily. In particular we note that all these functions are continuously differentiable on $\mathcal{U}[\tau_0, \tau_1]$ because integration is a continuous operation and a depends smoothly on H in $\mathcal{U}[\tau_0, \tau_1]$. Therefore it suffices to analyze the differentiability of the integral (8.8) appearing in the regularized two-point distribution. Moreover, within χ_k only the potential V is (smoothly on $\mathcal{U}[\tau_0, \tau_1]$) dependent on H, thus simplifying the computations considerably.[2] Continuing with the regularized two-point distribution order by order as in Prop. 8.1, we have:

1st order.   Since

$$\mathrm{d}\left(\chi_k^0 \overline{\chi}_k^1 + \chi_k^1 \overline{\chi}_k^0 + \frac{V}{4k_0^3}\right)(H; \delta H)(\tau)$$

$$= \frac{1}{4k_0^3}\int_{\tau_0}^{\tau} \cos\left(2k_0(\eta - \tau)\right)\mathrm{d}V'(H; \delta H)(\eta)\,\mathrm{d}\eta,$$

we can proceed with the proof as in Prop. 8.1 with V' replaced by $\mathrm{d}V'$ and differentiability follows.

2nd order.   As above, this part of the proof can be shown by replacing occurrences of V and V' in (8.12) of Prop. 8.1 with $\mathrm{d}V$ and $\mathrm{d}V'$ respectively.

---

[2]If we were to work in cosmological time as in [178], we would also have to consider the functional dependence of conformal time on the scale factor.

Higher orders.   For orders $n > 2$ we can again use an estimate similar to (6.20) to obtain a result analogous to (8.13):

$$\left| d\left(\sum_{n=3}^{\infty} \sum_{l=0}^{n} \chi_k^l \, \overline{\chi}_k^{n-l} \right) (H; \delta H)(\tau) \right| \le \frac{4}{k_0^4} \left(\int_{\tau_0}^{\tau} \left| dV(H; \delta H)(\eta) \right| d\eta \right)$$

$$\times \left(\int_{\tau_0}^{\tau} \left| V(\eta) \right| d\eta \right)^2 \exp\left(2 \int_{\tau_0}^{\tau} (\tau - \eta) \left| V(\eta) \right| d\eta \right).$$

In this way we can conclude the proof of the present proposition.   □

We can now formulate the main theorem of this chapter:

Theorem 8.5. *Let $(a_0, H_0), a_0 > 0, |H_0| < H_c$, be some initial conditions fixed at τ_0 for (8.5). There is a non-empty interval $[\tau_0, \tau_1]$ and a closed subset $U \subset C[\tau_0, \tau_1]$ on which a unique solution to (8.5) exists.*

Proof. In Prop. 8.4 we showed that f is continuously differentiable on $\mathcal{U}[\tau_0, \tau_1]$ for any τ_1. Using Prop. 3.11, we can thus find a $\tau_1 > \tau_0$ and a closed subset $U \subset \mathcal{U}[\tau_0, \tau_1]$ such that $F(U) \subset U$. It then follows from Prop. 3.10 that F has a unique fixed point in U.   □

Notice that the solution provided by the previous theorem is actually more regular, it is at least differentiable. Thus the corresponding spacetime is C^2 and has well-defined curvature tensors. The extra regularity is provided by (8.5) and can be easily seen when it is written in its differential form (8.4). It might be surprising that the solutions are not smooth, since the procedure to find the solution involves repeated integration, but because the chosen adiabatic state is only guaranteed to be continuous on every spacetime, H is only C^1. Using Cor. 3.6, one can see that a more regular state immediately improves also the regularity of the solution.

8.3   Global solutions

In this section we would like to show that it is always possible to extend a 'regular' local solution up to the point where either H^2 becomes bigger than H_c^2 or a diverges.[3] To this end we start giving a definition we shall use below.

Definition 8.6. *A continuous solution H_* of (8.5) in the interval $[\tau_0, \tau_1]$ with initial conditions*

$$a(H_*)(\tau_0) = a_0, \qquad H_*(\tau_0)^2 = H_0^2 = \frac{1}{3}\left(\rho(\tau_0) + \Lambda \right)$$

will be called regular*, if no singularity for either a, H_* or H_*' is encountered in $[\tau_0, \tau_1]$. Namely, H_* must satisfy the following conditions:*

1. $\left\| H_*(\tau) \right\|_{C[\tau_0, \tau_1]} < H_c,$

2. $a_0 \int_{\tau_0}^{\tau} H_*(\eta) \, d\eta < 1$ *for every τ in $[\tau_0, \tau_1]$.*

---

[3]$H^2 = H_c^2$ corresponds to a singularity in the derivative of H in (8.4).

We remark that a local solution obtained from Thm. 8.5 is a regular solution. Henceforth, assume that we have a regular solution H_* as described in the definition. Notice that condition a) ensures that no singularity in H'_* is met in $[\tau_0, \tau_1]$. Condition b), on the other hand, ensures that a does not diverge in the interval $[\tau_0, \tau_1]$. Moreover, both a) and b) together imply that a is strictly positive, as can be seen from (8.17).

We would like to prove that a regular solution can always be extended in $C[\tau_0, \tau_2]$ for a sufficiently small $\tau_2 - \tau_1 > 0$. To this end, let us again consider the set

$$\mathcal{U}[\tau_1, \tau_2] \doteq \left\{ H \in C[\tau_1, \tau_2] \;\middle|\; \|H\|_{C[\tau_1, \tau_2]} < \min\left\{ a_1^{-1}(\tau_2 - \tau_1)^{-1}, H_c \right\} \right\}$$

defined in (8.18) and where $a_1 \doteq a(H_*)(\tau_1)$ is the value assumed by the solution $a(H_*)$ at τ_1. Now we can give a proposition similar to Prop. 8.4, namely:

Proposition 8.7. *Let H_* be a solution of (8.5) in $C[\tau_0, \tau_1]$ which is also regular. The functional $f(H)$ of (8.19), when evaluated on regular extensions of H_* in $\mathcal{U}[\tau_1, \tau_2]$, is continuously differentiable for arbitrary $\tau_2 > \tau_1$.*

Proof. The proof of this proposition can be obtained exactly as the proof of Prop. 8.4. However, the estimates we have obtained in Prop. 8.2 and the proof of Prop. 8.4 cannot be applied straightforwardly because the state ω depends on the initial time τ_0 and the initial datum a_0 through the construction described in Sect. 6.2.4. Moreover, the estimates of Prop. 8.2 depend on the knowledge of a and a' on the whole interval $[\tau_0, \tau_2]$. Luckily enough, we know that the solution H_* is regular in $[\tau_0, \tau_1]$, while we know that the extension restricted to $[\tau_1, \tau_2]$ is in the set $\mathcal{U}[\tau_1, \tau_2]$; thus we just need to use the following estimates

$$\|a\|_{C[\tau_0, \tau_2]} = \max\left\{ \|a\|_{C[\tau_1, \tau_2]}, \|a\|_{C[\tau_0, \tau_1]} \right\},$$
$$\|a\|_{C[\tau_0, \tau_2]}^{-1} = \max\left\{ \|a\|_{[\tau_1, \tau_2]}^{-1}, \|a\|_{C[\tau_0, \tau_1]}^{-1} \right\},$$
$$\|a'\|_{C[\tau_0, \tau_2]} = \max\left\{ \|a'\|_{C[\tau_1, \tau_2]}, \|a'\|_{C[\tau_0, \tau_1]} \right\}.$$

With this in mind, we can again use Prop. 8.2 to control the boundedness of $\omega(:\widehat{\varphi}^2:)$. Then, making the replacements $\tau_0 \to \tau_1, \tau_1 \to \tau_2$ and $a_0 \to a_1$ at the *appropriate* places in Prop. 8.4, one can see that estimates are not substantially influenced and that thesis still holds for $\mathcal{U}[\tau_1, \tau_2]$. □

Notice that it is always possible to fix τ_2 such that $a_1^{-1}(\tau_2 - \tau_1)^{-1} \geq H_c$, whereby $\mathcal{U}[\tau_1, \tau_2]$ becomes the set of *all* possible regular extensions of H_* in $[\tau_1, \tau_2]$. This guarantees that any extension in $\mathcal{U}[\tau_1, \tau_2]$ is the unique regular extension.

We are now ready to state the main theorem of the present section which can be proven exactly as Thm. 8.5.

Theorem 8.8. *Consider a solution $H_*(\tau)$ in $C[\tau_0, \tau_1]$ of (8.5). If the solution is regular in $[\tau_0, \tau_1]$, as defined in Def. 8.6, then it is possible to find a $\tau_2 > \tau_1$ such that, the solution H_* can be extended uniquely to $C[\tau_0, \tau_2]$ and the solution is regular therein.*

Proof. Thanks to Prop. 8.7, f is continuously differentiable on all regular extensions of H_* in $\mathcal{U}[\tau_1, \tau_2]$ for any τ_2 such that $a_1^{-1}(\tau_2 - \tau_1)^{-1} \geq H_c$. With the remarks of

the proof of Prop. 8.7 we can use Prop. 8.2 to estimate the boundedness of $\omega(:\widehat{\varphi}^2:)$ and apply Prop. 3.11 to find a $\tau_2 > \tau_1$ and a closed subset $U \subset \mathcal{U}[\tau_1, \tau_2]$ such that $F(U) \subset U$. It then follows from Prop. 3.10 that F has a unique fixed point in U.   □

We study now all possible solutions of (8.5) which are defined on intervals of the form $[\tau_0, \tau)$, which are regular on any closed interval contained in their domain and which enjoy the same initial values $a_0 = a(\tau_0), H_0 = H(\tau_0)$.

Proposition 8.9. *A maximal solutions exists; it is unique and regular.*

Proof. Let $\mathcal{S} = \{I_\alpha, H_\alpha\}_{\alpha \in A}$, with $A \subset \mathbb{N}$ some index set, be the set of all possible regular solutions H_α with domain I_α for the same initial values. By the existence of local solutions $\mathcal{S}$ is not empty. We then take the union $I = \bigcup_{\alpha \in A} I_\alpha$ and define $H(\tau) = H_\alpha(\tau)$ for $\tau \in I_\alpha$, which is a well-defined regular solution by Prop. 8.7. Since every I is a superset of every I_α, H is the unique maximal regular solution.   □

As for the solution provided by theorem 8.5, also the maximal solution obtained above correspond to a metric with C^2 regularity.

8.4   Numerical solutions

The first problem that one encounters when attempting to treat the semiclassical Einstein equation in a numerical fashion, is the construction of states. Here, the mode equation (6.10), which describes an oscillator with a time-dependent 'resonance frequency' $\omega_k^2 = k^2 + a^2 m^2$, has to be solved. Standard numerical solvers, like the Runge–Kutta method, rely on differentiation and their error scales like a (high-order) derivative of the solution. However, each derivative of an oscillating function increases the amplitude by a power of the frequency, thus ultimately leading to large errors for quickly oscillating differential equations after a short time span. This problem can be partially counteracted by choosing ever smaller step sizes in time, but eventually one will encounter a computational barrier. Another possibility is to look for a non-standard approach to solve the mode equation. Such methods replace differentiation with integration, but, since the numerical integration of highly oscillatory functions is also a non-trivial problem, this is still an active area of research, see for example [84, 131, 132, 143, 144].

One might say, that the large frequency behaviour of the modes is of no relevance when solving the semiclassical Einstein equation because it involves the state only *after* regularization, *i.e.*, after the terms that contribute for large ω_k have been subtracted. While this response is, to some degree, certainly true, numerical errors in the solution before the regularization and in the subtraction itself can still accumulate. Therefore, the issue of the high frequency modes and their regularization has to be carefully addressed in a numeric approach.

Although the perturbative construction of the state used in this chapter (see also Sect. 6.2.4), the functional (8.5) and the use of the Banach fixed-point theorem in the proof of Thm. 8.5 were not developed with a numerical application in mind, there are reasons why they might be useful also for numerics: The mode solution are found recursively from (6.19), an integral equation which avoids the differentiation problem

discussed above. Moreover, as seen in Prop. 8.1, only the first two partial modes $\chi_k^{(0)}$ and $\chi_k^{(1)}$ are affected by the regularization in a well-understood way, so that also the numerical difficulty in the regularization can circumvented. Nevertheless, also this approach is not without its problems as it involves repeated numerical integration of oscillating functions and therefore it is very slow if naïvely implemented as a Riemann sum, because it requires small time steps. The other feature of the proof of existence that allows a translation to numerics is the use of the Banach fixed-point theorem. Namely, we can be assured that an iterated recursive application of (8.5) will converge to a solution, even though we do not know how quickly convergence occurs.

8.5 Outlook

In this chapter we have studied the backreaction of a quantum massive scalar field conformally coupled with gravity to cosmological spacetimes. We have given initial conditions at finite time $\tau = \tau_0$ and we have shown that a unique maximal solution exists. The maximal solution either lasts forever or until a spacetime singularity is reached.

In order to obtain this result, we have used a state which looks as much as possible like the vacuum at the initial time. Notice that it is possible to choose other classes of states without significantly altering the results obtained in this chapter. In particular, if we restrict ourself to Gaussians pure state which are homogeneous and isotropic, their two-point function takes the form

$$\widetilde{\omega}_2(x, y) = \lim_{\varepsilon \to 0^+} \frac{1}{(2\pi)^3} \int_{\mathbb{R}^3} \frac{\overline{\xi}_k(\tau_x)}{a(\tau_x)} \frac{\xi_k(\tau_y)}{a(\tau_y)} \, e^{i\vec{k}\cdot(\vec{x}-\vec{y})} e^{-\varepsilon k} \, d\vec{k},$$

where ξ_k are solutions of (6.10) which enjoy the Wronskian condition (6.11). These χ_k can then be written as a Bogoliubov transformation of the modes χ_k studied earlier in this chapter, namely,

$$\xi_k = A(k)\chi_k + \overline{B}(k)\overline{\chi}_k$$

for suitable functions A and B. Then, because of the constraint $|A|^2 - |B|^2 = 1$, the difference

$$\widetilde{\omega}(:\widehat{\varphi}^2:) - \omega(:\widehat{\varphi}^2:) = \lim_{\varepsilon \to 0^+} \frac{1}{(2\pi)^3} \frac{2}{a^2} \int_{\mathbb{R}^3} \left(|B|^2 \chi_k \overline{\chi}_k + \mathrm{Re}\left(AB\chi_k\chi_k \right) \right) e^{-\varepsilon k} \, d\vec{k}$$

can be easily controlled employing (6.20) if $|B|$ is sufficiently regular (*e.g.*, if $B(k)$ is in $L^2 \cap L^1$).[4] With this observation it is possible to obtain again all the estimates used in the proofs of Thms. 8.5 and 8.8.

In the future, it would be desirable to study the semiclassical equations in more general cases, namely for more general fields, abandoning for example the conformal coupling, and for more general background geometries. The results presented here cannot straightforwardly be extended to fields which are not conformally coupled to curvature or to spacetimes that are not conformally flat because in that case

---

[4]A detailed analysis of this problem is present in [221].

fourth order derivatives of the metric originating in the conformal anomaly cannot be cancelled by a judicious choice of renormalization parameters, *i.e.*, Wald's fifth axiom [218] cannot be satisfied. To still solve the semiclassical Einstein equation with methods similar to those presented here, a deeper analysis of the states is required, in particular, one needs states of higher regularity. A preliminary study in this direction can be found in a paper of Eltzner and Gottschalk [83], where the semiclassical Einstein equation on a FLRW background with non-conformally coupled scalar field is discussed. The case of backgrounds which are only spherically symmetric is interesting from many perspectives. Its analysis could give new hints on the problem of semiclassical black hole evaporation and confirm the nice two-dimensional results obtained in [19]. Finally, the limit of validity of the employed equation needs to be carefully addressed in the future.

9

Induced semiclassical fluctuations

Introduction

As described in Chap. 7, in semiclassical Einstein gravity one equates a classical quantity, the Einstein tensor, with the expectation value of a quantum observable, the quantum stress-energy tensor, *i.e.*, a quantity with a probabilistic interpretation. Such a system could make sense only when the fluctuations of the quantum stress-energy tensor can be neglected. Unfortunately, as also noticed in [178], the variance of quantum unsmeared stress-energy tensor is always divergent even when proper regularization methods are considered. The situation is slightly better when a smeared stress-energy tensor is analyzed. In that way, however, the covariance of (7.1) gets lost. A possible way out is to allow for fluctuations also on the left-hand side of (7.1). This is the point of view we shall assume within this chapter, which is based on an article [4] in collaboration with Pinamonti.

More precisely, we interpret the Einstein tensor as a stochastic field and equate its n-point distributions with the symmetrized n-point distributions of the quantum stress-energy tensor. As an application of this (toy) model, we analyze the metric fluctuations induced by a massive, conformally coupled scalar field via the (quantum) stress-energy tensor in the simplest non-trivial spacetime – de Sitter spacetime. We find that the potential in a Newtonianly perturbed FLRW spacetime has a almost scale-invariant power spectrum.

These results encourage a comparison with the observation of anisotropies in the cosmic microwave background and their theoretical explanations. Anisotropies in the angular temperature distribution were predicted by Sachs and Wolfe [188] shortly after the discovery of the cosmic microwave background (CMB) by Penzias and Wilson [174]. In their famous paper they discuss what was later coined the *Sachs–Wolfe effect*: The redshift in the microwave radiation caused by fluctuations in the gravitational field and the corresponding matter density fluctuations. In the standard model of inflationary cosmology the fluctuations imprinted upon the CMB are seeded by quantum fluctuations during inflation [161, 162], see also the reviews in [76, 80].

The usual computation of the power spectrum of the initial fluctuations produced by single-field inflation can be sketched as follows [26, 76, 80]: First, one introduces a (perturbed) classical scalar field $\varphi + \delta\varphi$, the inflaton field, which is coupled to a (perturbed) expanding spacetime $g + \delta g$. Then, taking the Einstein equation and the Klein–Gordon equation at first order in the perturbation variables, one constructs an equation of motion for the *Mukhanov–Sasaki variable* $Q = \delta\varphi + \dot{\varphi} H^{-1}\Phi$, where Φ is the Bardeen potential [25] and H the Hubble constant. Q is then quantized[1] (in the slow-roll approximation) and one chooses as the state of the associated quantum field

---

[1] A recent discussion about the quantization of a such system can be found in [81].

a Bunch–Davies-like state. Last, one evaluates the power spectrum $P_Q(k)$ of Q, *i.e.*, the Fourier-transformed two-point distribution of the quantum state, in the super-Hubble regime $k \ll aH$ and obtains an expression of the form[2]

$$P_Q(k) = \frac{A_Q}{k^3} \left(\frac{k}{k_0} \right)^{n_s - 1}, \tag{9.1}$$

where A_Q is the amplitude of the fluctuations, k_0 a pivot scale and n_s the spectral index. Notice the factor of k^{-3} in (9.1) which gives the spectrum the 'scale-invariant' *Harrison–Zel'dovich* form if $n_s = 1$. Depending on the details of model, $n_s \lesssim 1$ and there is also a possibility for a scale dependence of n_s – the 'running' of the spectral index $n_s = n_s(k)$.

This result can then be related to the power spectrum of the comoving curvature perturbation $\mathcal{R}$, which is proportional to Q, and can be compared with observational data. Assuming adiabatic and Gaussian initial perturbations, the WMAP collaboration finds $n_s = 0.9608 \pm 0.0080$ (at $k_0 = 0.002\,\mathrm{Mpc}^{-1}$) in a model without running spectral index and gravitational waves, excluding a scale-invariant spectrum at 5σ [121]. Furthermore, the data of WMAP and other experiments can be used to constrain the deviations from a pure Gaussian spectrum, the so called non-Gaussianities, that arise in some inflationary models [26, 33, 150].

In [7–9, 170] concerns have been raised whether the calculation leading to (9.1) and similar calculations are correct: The authors argue that the two-point distribution of the curvature fluctuations has to be regularized and renormalized similarly to what is done in semiclassical gravity. As a result the power spectrum is changed sufficiently that previously observationally excluded inflation models become realistic again. On the contrary the authors of [77, 153] argue that the adiabatic regularization employed in [7–9, 170] is not appropriate for low momentum modes if evaluated at the Horizon crossing and irrelevant for these modes if evaluated at the end of inflation.

A slightly different approach to the calculation of the power spectrum based on *stochastic gravity* can be found in [129, 184, 185]. In spirit similar to the approach presented in this chapter, the authors equate fluctuations of the stress-energy tensor with the correlation function of the Bardeen potential. In the super-Hubble regime they obtain an almost scale-invariant power spectrum. Moreover, they discuss the equivalence of their stochastic gravity approach with the usual approach of quantizing metric perturbations.

Our approach here is strictly different from the standard one described above. Instead of quantizing a coupled system of linear inflaton and gravitational perturbations, we aim at extending the semiclassical Einstein equation to describe metric fluctuations via the fluctuations in the stress-energy tensor of a quantum field.

9.1   Fluctuations of the Einstein tensor

Consider now the Einstein tensor as a random field. Then we could imagine to equate the probability distribution of the Einstein tensor with the probability distribution of

---

[2]An alternative definition of the power spectrum is $\mathscr{P}_Q(k) = (2\pi)^{-2} k^3 P_Q(k)$.

the stress-energy tensor. This suggestion, however, seems largely void without a possibility of actually computing the probability distributions of the stress-energy tensor because, as discussed above, its moments of order larger than one are divergent.

Instead we may approach this idea by equating the hierarchy of n-point distributions of the Einstein tensor with that of the stress-energy tensor:

$$\langle G_{ab}(x_1)\rangle = \omega(:T_{ab}(x_1):), \tag{9.2a}$$

$$\langle \delta G_{ab}(x_1)\,\delta G_{c'd'}(x_2)\rangle = \tfrac{1}{2}\omega(:\delta T_{ab}(x_1): :\delta T_{c'd'}(x_2): + :\delta T_{c'd'}(x_2): :\delta T_{ab}(x_1):), \tag{9.2b}$$

and

$$\langle \delta G^{\boxtimes n}\rangle = \omega\big(\operatorname{Sym}(:\delta T:^{\boxtimes n})\big), \quad n > 1, \tag{9.2c}$$

where ω is a Hadamard state and we defined

$$\delta G_{ab} \doteq G_{ab} - \langle G_{ab}\rangle \quad\text{and}\quad :\delta T_{ab}: \doteq :T_{ab}: - \omega(:T_{ab}:).$$

The symmetrization on the right-hand side is necessary because the classical quantity on left-hand side is invariant under permutation.

We emphasize that we are equating singular objects in (9.2c). Having all the n-point distributions of the Einstein stochastic tensor, we can easily construct an equation for the moments of the smeared Einstein tensor which equals the moments of a smeared stress-energy tensor by smearing both sides of (9.2) with tensor products of a smooth compactly supported function. This smearing also automatically accounts for the symmetrization in (9.2).

Furthermore we stress that equating moments, obtained smearing both side of (9.2), is not equivalent to equating probability distributions. Although it is also possible to arrive at a description in terms of moments when coming from a probability distribution, the inverse mapping is not necessarily well-defined. Successful attempts to construct a probability distribution for smeared stress-energy tensors can be found in [90, 91].

Consider now a quasi-free Hadamard state ω of a conformally coupled scalar field φ on a spacetime $(M, \overline{g})$, the *background spacetime*. Our aim is to calculate the perturbation of the background spacetime as specified by the correlation functions on the left-hand side of (9.2) due to the fluctuations of the stress-energy in the quantum state ω as specified on the right-hand side of (9.2). In particular we will require that ω satisfies (9.2a) when we identify the Einstein tensor of the background spacetime $\overline{G}_{ab}$ with $\langle G_{ab}\rangle$ (*cf.* Chap. 8 for a discussion of the solutions of the semiclassical Einstein equation in cosmological spacetimes). Note that by choosing this Ansatz we are completely ignoring any backreaction effects of the fluctuations to the background metric and evaluate the stress-energy tensor on a state specified on the background spacetime.

Later on we consider perturbations of the scalar curvature induced by a 'Newtonianly' perturbed FLRW metric. For this reason it will be sufficient to work with the trace of (9.2) (using the background metric) instead of the full equations. With the definition

$$S \doteq -\overline{g}^{ab} G_{ab},$$
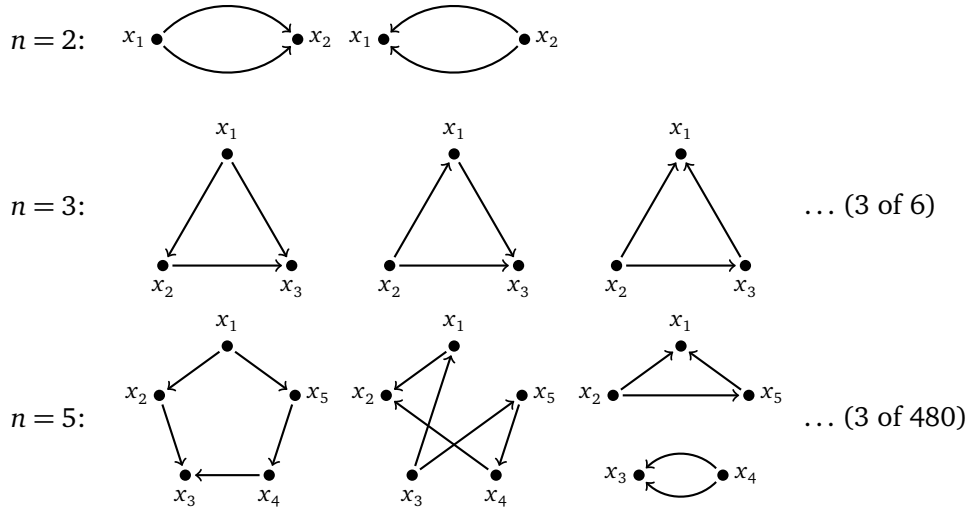
Figure 9.1. A few graphs illustrating (9.3c) for $n = 2$, $n = 3$ and $n = 5$.

such that $R = \langle S \rangle$, the equations (9.2) simplify to

$$\langle S(x_1) \rangle = \frac{m^2}{8\pi^2}[w] - \frac{1}{4\pi^2}[v_1] + \text{ren. freedom}, \quad (9.3a)$$

$$\langle S(x_1)S(x_2) \rangle - \langle S(x_1) \rangle \langle S(x_2) \rangle = m^4 \big(\omega_2^2(x_1, x_2) + \omega_2^2(x_2, x_1) \big), \quad (9.3b)$$

and

$$\langle (S - \langle S \rangle)^{\boxtimes n}(x_1, \ldots, x_n) \rangle = 2^n m^{2n} \operatorname{Sym} \left(\sum_\Gamma \prod_{i,j} \frac{\omega_2^{\lambda_{ij}^\Gamma}(x_i, x_j)}{\lambda_{ij}^\Gamma!} \right), \quad (9.3c)$$

where the sum is over all directed graphs Γ with n vertices $1, \ldots, n$ with two arrows at every vertex directed to a vertex with a larger label. $\lambda_{ij}^\Gamma \in \{0, 1, 2\}$ is the number of arrows from i to j. If we perform the symmetrization in (9.3c), we see that the sum is over all acyclical directed graphs with two arrows at every vertex. For illustration some graphs are shown in Fig. 9.1.

To obtain (9.3b) and (9.3c), note that $:\widehat{\varphi}^2: - \omega(:\widehat{\varphi}^2:)$ does not depend on the choice of normal ordering[3] and thus only (9.3a) needs to be renormalized. Therefore we may choose normal ordering with respect to ω_2 to see that the combinatorics are equivalent to those in Minkowski space. Moreover, as ω_2 is a bisolution of the Klein–Gordon equation, the term $\frac{1}{3}:\widehat{\varphi} \operatorname{P}\widehat{\varphi}:$ which causes the trace anomaly in (9.3a) (cf. Sect. 7.2.4) does not contribute to the higher moments.

9.2   Fluctuations around a de Sitter spacetime

We shall now specialize the general discussion presented above to Newtonianly perturbed, exponentially expanding, flat FLRW universes. That is, the background

---

[3]Indeed this holds true if we replace φ^2 with $L\varphi^2$, for any linear operator L.

spacetime $(M, \overline{g})$ is given in conformal time $\tau < 0$ by the metric tensor

$$\overline{g} \doteq (H\tau)^{-2}(-\mathrm{d}\tau \otimes \mathrm{d}\tau + \delta_{ij}\,\mathrm{d}x^i \otimes \mathrm{d}x^j)$$

and we consider fluctuations of the scalar curvature derived from metric perturbations of the form

$$g \doteq (H\tau)^{-2}\big(-(1+2\Phi)\,\mathrm{d}\tau \otimes \mathrm{d}\tau + (1-2\Phi)\,\delta_{ij}\,\mathrm{d}x^i \otimes \mathrm{d}x^j\big). \tag{9.4}$$

The kind of fluctuations that we consider by choosing (9.4) resemble those that are present in single-scalar field inflation in the longitudinal gauge, where there are only 'scalar fluctuations' without anisotropic stress (so that the two Bardeen potentials coincide) [80, 162]. Notice that, for classical metric perturbation, these constraints descend from the linearized Einstein equation, however, *a priori* there is no similar constraint in (9.2b). Despite these facts, we proceed analyzing the influence of quantum matter on this special kind of metric perturbations and we also refrain from discussing the gauge problem associated to choosing a perturbed spactime; the chosen perturbation potential Φ is *not* gauge invariant.

We can now calculate the various perturbed curvature tensors and obtain in particular

$$S = 12H^2(1-3\Phi) + 24H^2\tau\partial_\tau\Phi - 6H^2\tau^2\partial_\tau^2\Phi + 2H^2\tau^2\,\vec{\nabla}^2\Phi + O(\Phi^2)$$

for the trace of the perturbed Einstein tensor, where $\vec{\nabla}^2 = \partial_{x^1}^2 + \partial_{x^2}^2 + \partial_{x^3}^2$ is the ordinary Laplace operator. Dropping terms of higher than linear order, this can also be written as

$$S - \langle S \rangle = -6H^2\tau^4\big(\partial_\tau - \tfrac{1}{3}\vec{\nabla}^2\big)\tau^{-2}\Phi, \tag{9.5}$$

where $\langle S \rangle = 12H^2$ is nothing but the scalar curvature of the background spacetime. Notice that, up to a rescaling, the operator on the right-hand side of (9.5) looks like a wave operator with the characteristic velocity equal to $1/\sqrt{3}$ of the velocity of light.

We can now evaluate the influence of quantum matter fluctuations on the metric fluctuations by inverting the previous hyperbolic operator by means of its retarded fundamental solutions Δ_{ret} and applying it on both sides of (9.3b) and (9.3c). From (9.3b) we can then (formally) obtain the two-point correlation functions of Φ (per definition $\langle \Phi \rangle = 0$):

$$\langle \Phi(x_1)\Phi(x_2) \rangle = m^4 \iint_{\mathbb{R}^8} \Delta_{\mathrm{ret}}(x_1, y_1)\Delta_{\mathrm{ret}}(x_2, y_2)\big(\omega_2^2(y_1, y_2) \\ + \omega_2^2(y_2, y_1)\big)\,\mathrm{d}^4 y_1\,\mathrm{d}^4 y_2. \tag{9.6}$$

Employing the retarded fundamental solutions in the inversion without adding any solution of (9.5), we are implicitly assuming that all the n-point distributions of the perturbation potential Φ are sourced by quantum fluctuations. Here we are only interested in evaluating their effect.

9.2.1 The squared two-point distribution

In order to proceed with our analysis, we shall specify the quantum state ω for the matter theory. Following the Ansatz discussed in the preceding section, we choose a

quasi-free Hadamard state which satisfies the semiclassical Einstein equation on the background. In particular, we require that ω solves (9.3a), namely

$$12H^2 = \frac{m^2}{8\pi^2}[w] - \frac{1}{4\pi^2}[v_1] + 4c_1 m^4 - c_2 m^2 R$$

The right-hand side of the previous equation is characterized by three contributions: The state dependent part $[w]$, the anomaly part $[v_1]$, which takes the simple form (cf. Eq. (7.9))

$$[v_1] = -\frac{H^4}{60} + \frac{m^4}{8},$$

and the renormalization freedom $c_1 m^4$ and $c_2 m^2 R$. Here we set $c_2 = 0$, because we assume the point of view that we have already measured Newton's gravitational constant and do not wish to renormalize it (cf. Sect. 7.2.3). That is, we have

$$12H^2 = \frac{m^2}{8\pi^2}[w] + \frac{1}{8\pi^2}\left(\frac{H^4}{15} - \frac{m^4}{2}\right) + 4c_1 m^4.$$

For the semiclassical Einstein equation to hold, we therefore have to require that $[w]$ is a constant. Then, having fixed H and m (no matter their absolute value), there is always a choice of c_1 for which the chosen metric $\overline{g}$ and ω satisfy the semiclassical Einstein equation. On a de Sitter spacetime these criteria are satisfied by the Bunch–Davis state, cf. Sect. 6.2.1.

In order to evaluate the influence of the quantum matter fluctuations on Φ via equation (9.6), we have to discuss the form of the two-point distribution of the chosen state and its square. Any Hadamard state on $(M, \overline{g})$ can be written is equal to the Bunch–Davies state up to smooth terms. In particular, the two-point distribution ω_2 of every Hadamard state on de Sitter spacetime is of the form

$$\omega_2(x, x') = \lim_{\varepsilon \to 0^+} \frac{H^2}{4\pi^2} \frac{\tau\tau'}{(x - x')^2 + 2i\varepsilon(\tau - \tau') + \varepsilon^2} + \text{less singular terms}, \quad (9.7)$$

where we write $(x - x')^2 \doteq -(\tau - \tau')^2 + (\vec{x} - \vec{x}')^2$ and, as always, the limit $\varepsilon \to 0^+$ is a weak limit. It is no surprise that the leading singularity is conformally related to the two-point distribution of a massless scalar field on Minkowski spacetime, we denote it by $\omega_{\mathbb{M}}$. Thus it is also clear that the less singular contributions vanish in the limit of zero mass.

As can be seen in (9.3b), we need to compute the square of the two-point distribution of the state in question. For our purposes it will be sufficient to compute the square of the leading singularity in the Hadamard state.[4] The square of the massless two-point distribution on Minkowski space is

$$\omega_{\mathbb{M}}(x, x')^2 = \lim_{\varepsilon \to 0^+} \left(\frac{1}{4\pi^2} \frac{1}{(x - x')^2 + 2i\varepsilon(\tau - \tau') + \varepsilon^2}\right)^2.$$

---

[4]Note that, in (spatial) momentum space, the leading singularity in (9.7) contributes the smallest inverse power of the momentum $\vec{k}$; the 'less singular terms' correspond to higher inverse powers of $\vec{k}$. Accordingly, these terms fall off faster for large $\vec{k}$. This is nothing but the usual relation ship between high momenta and short distances.

Writing $\omega_{\mathbb{M}}$ in terms of its spatial Fourier transform, an expression for the spatial Fourier transform of the square of the massless Minkowski vacuum can be obtained as

$$\omega_{\mathbb{M}}(x,x')^2 = \lim_{\varepsilon \to 0^+} \frac{1}{128\pi^5} \int_{\mathbb{R}^3} e^{i\vec{k}\cdot(\vec{x}-\vec{x}')} \int_k^{\infty} e^{-ip(\tau-\tau')} e^{-\varepsilon p}\, dp\, d\vec{k}. \tag{9.8}$$

Later on we will use this expression in order to obtain the power spectrum of Φ.

9.2.2 Power spectrum of the metric perturbations

We want to compute the power spectrum $P(\tau, \vec{k})$ of the two-point correlation of Φ at the time τ. Since both the spacetime and the chosen state are invariant under spatial translation, it can be defined as

$$\langle \Phi(\tau,\vec{x})\Phi(\tau,\vec{x}')\rangle \doteq \frac{1}{(2\pi)^3} \int_{\mathbb{R}^3} P(\tau,\vec{k})\, e^{i\vec{k}\cdot(\vec{x}-\vec{x}')}\, d\vec{k}.$$

To obtain P, we first need an expression for the retarded operator Δ_{ret} corresponding to (9.5):

$$(\Delta_{\mathrm{ret}} f)(\tau,\vec{x}) = \frac{1}{(2\pi)^3} \int_{\mathbb{R}^3}\int_{-\infty}^{\tau} \widehat{\Delta}_{\mathrm{ret}}(\tau,\tau_1,\vec{k})\widehat{f}(\tau_1,\vec{k})e^{i\vec{k}\cdot\vec{x}}\, d\tau_1\, d\vec{k}, \quad \text{with}$$

$$\widehat{\Delta}_{\mathrm{ret}}(\tau,\tau_1,\vec{k}) \doteq -\frac{1}{6H^2}\frac{\tau^2}{\tau_1^4}\frac{\sqrt{3}}{k}\sin\left(k\,(\tau-\tau_1)/\sqrt{3}\right),$$

where f is a compactly supported smooth function. We can then rewrite (9.6) in Fourier space to obtain

$$P(\tau,\vec{k}) = 2m^4 \int_{-\infty}^{\tau}\int_{-\infty}^{\tau} \widehat{\Delta}_{\mathrm{ret}}(\tau,\tau,\vec{k})\widehat{\Delta}_{\mathrm{ret}}(\tau,\tau',\vec{k})\widehat{\omega_{BD}^2}(\tau,\tau',\vec{k})\, d\tau\, d\tau'.$$

Note that the symmetrization of the state is taken care of indirectly by the equal limits of the two integrations.

As discussed above (see (9.7) and the following paragraph), we will compute the contribution due to the leading singularity of the Hadamard state:

$$P_0(\tau,\vec{k}) \doteq 2H^4 m^4 \int_{-\infty}^{\tau}\int_{-\infty}^{\tau} \widehat{\Delta}_{\mathrm{ret}}(\tau,\tau,\vec{k})\,\widehat{\Delta}_{\mathrm{ret}}(\tau,\tau',\vec{k})\,\tau^2\tau'^2\widehat{\omega_{\mathbb{M}}^2}(\tau_1,\tau',\vec{k})\, d\tau\, d\tau'.$$

We emphasize at this point that, because of the form of (9.8) and of $\widehat{\Delta}_{\mathrm{ret}}$, no infrared (with respect to $\vec{k}$) singularity appears in $P_0(\tau,\vec{k})$ at finite τ. Recall also that the error we are committing, using $P_0(\tau,\vec{k})$ at the place of $P(\tau,\vec{k})$, tends to vanish in the limit of small masses. Inserting the spectrum of $\omega_{\mathbb{M}}^2$ obtained in (9.8) and switching the order in which the integrals are taken (for $\varepsilon > 0$), we can write

$$P_0(\tau,\vec{k}) = \lim_{\varepsilon \to 0^+} \frac{m^4}{16\pi^2} \int_k^{\infty} \frac{1}{k^4}\left|A\left(\tau,k/\sqrt{3},p\right)\right|^2 e^{-\varepsilon p}\, dp, \tag{9.9}$$

where we have introduced the auxiliary function

$$A(\tau,\kappa,p) \doteq \int_{-\infty}^{\tau} \frac{\kappa\,\tau^2}{\tau_1^2}\sin\left(\kappa\,(\tau-\tau_1)\right) e^{-ip\tau_1}\, d\tau_1, \tag{9.10}$$

which can also be written in closed form in terms of the generalized exponential integral[5] E_2 as

$$A(\tau, \kappa, p) = A(\kappa\tau, p\tau) = \frac{i}{2}\kappa\tau \left(E_2\big(i(p+\kappa)\tau\big)\,e^{i\kappa\tau} - E_2\big(i(p-\kappa)\tau\big)\,e^{-i\kappa\tau} \right)$$

for $p \geq \kappa > 0$ and by the complex conjugate of this expression if $\kappa > p, \kappa > 0$. In the following study of the form of the power spectrum P_0 the auxiliary function A will be instrumental.

Lemma 9.1. *For $|p| \neq \kappa > 0$, $A(\tau, \kappa, p)$ has the τ-uniform bound*

$$|A| \leq \frac{4\kappa^2}{|\kappa^2 - p^2|}. \tag{9.11}$$

For large negative times it satisfies the limit

$$\lim_{\tau \to -\infty} |A| = \frac{\kappa^2}{|\kappa^2 - p^2|}. \tag{9.12}$$

Proof. Using the fact that

$$e^{-ip\tau_1} = \left(\frac{d^2}{d\tau_1^{\,2}} + \kappa^2 \right) \frac{e^{-ip\tau_1}}{\kappa^2 - p^2},$$

we can perform two integrations by parts to obtain

$$A(\tau, \kappa, p) = \frac{\kappa^2}{\kappa^2 - p^2} \left(e^{-ip\tau} + R(\tau, \kappa, p) \right), \quad \text{with}$$

$$R(\tau, \kappa, p) \doteq \tau^2 \int_{-\infty}^{\tau} \left(\frac{4}{\tau_1^3} \cos\big(\kappa(\tau - \tau_1)\big) + \frac{6}{\kappa\tau_1^4} \sin\big(\kappa(\tau - \tau_1)\big) \right) e^{-ip\tau_1}\, d\tau_1.$$

It is now easy to obtain an upper bound for R which is uniform in conformal time, namely $|R| \leq 3$, which then yields the bound (9.11).

For the second part of the proposition we perform a change of the integration variable to $x = \tau_1/\tau$:

$$R(\tau, \kappa, p) = -\int_1^{\infty} \left(\frac{4}{x^3} \cos\big(\kappa\tau(1-x)\big) + \frac{6}{\kappa\,\tau\,x^4} \sin\big(\kappa\tau(1-x)\big) \right) e^{-ip\tau x}\, dx.$$

The contribution proportional to $1/\tau$ in R is bounded by $C(\kappa)/|\tau|$ and thus vanishes in the limit $\tau \to -\infty$. Moreover, since $|p| \neq \kappa$ and $1/x^3$ is L^1 on $[1, \infty)$, we can apply the Riemann–Lebesgue lemma and see that this contribution vanishes in the limit $\tau \to -\infty$. The remaining part of $|A|$ is $\kappa^2|\kappa^2 - p^2|^{-1}$, which is independent of τ, and thus the limit (9.12) holds true. □

Note that the bound for A obtained above is not optimal. Numerical integration indicates that $|A|^2$ is monotonically decreasing in τ and thus bounded by the limit stated in (9.12) (see also Fig. 9.2). Nevertheless, we can use this lemma to derive the following bounds and limits for P_0:

---

[5]For a definition and various properties of these special functions see *e.g.* [166, Chap. 8].

Proposition 9.2. *The leading contribution P_0 to the power spectrum of the potential Φ induced by a conformally coupled massive scalar field in the Bunch–Davis state is bounded by the Harrison–Zel'dovich spectrum uniformly in time, namely*

$$\left|P_0(\tau,\vec{k})\right| \le \frac{16\,C}{|\vec{k}|^3}, \quad C \doteq \frac{3 - 2\sqrt{3}\,\mathrm{arccoth}\,\sqrt{3}}{192\pi^2}\,m^4,$$

and it tends to the Harrison–Zel'dovich spectrum for $\tau \to -\infty$, i.e.,

$$\lim_{\tau \to -\infty} P_0(\tau,\vec{k}) = \frac{C}{|\vec{k}|^3}.$$

Proof. The proof can be easily obtained using the τ-uniform estimate (9.11) obtained in Lem. 9.1 and computing the integral

$$\left|P_0(\tau,\vec{k})\right| \le \frac{m^4}{\pi^2} \int_k^\infty \left(\frac{1}{3p^2 - k^2}\right)^2 \mathrm{d}p = \frac{3 - 2\sqrt{3}\,\mathrm{arccoth}\,\sqrt{3}}{12\pi^2}\frac{m^4}{k^3}.$$

Having shown the first part of the proposition, let us now analyze the limit

$$\lim_{\tau \to -\infty} P_0(\tau,\vec{k}) = \frac{m^4}{16\pi^2} \int_k^\infty \frac{1}{k^4}\lim_{\tau \to -\infty}\left|A\left(\tau, k/\sqrt{3}, p\right)\right|^2 \mathrm{d}p,$$

where we have taken the τ-limit before the integral and already evaluated the ε-limit because $|A|^2$ is bounded by an integrable function uniformly in time. Inserting the limit (9.12) from Lem. 9.1, we can compute the p-integral

$$\lim_{\tau \to -\infty} P_0(\tau,\vec{k}) = \frac{m^4}{16\pi^2} \int_k^\infty (3p^2 - k^2)^{-2}\,\mathrm{d}p = \frac{3 - 2\sqrt{3}\,\mathrm{arccoth}\,\sqrt{3}}{192\pi^2}\frac{m^4}{k^3},$$

thus concluding the proof. □

We can complement the results of Prop. 9.2 with the following observation:

Proposition 9.3. *The power spectrum P_0 has the form*

$$P_0(\tau,\vec{k}) = \frac{\mathscr{P}_0(|\vec{k}|\tau)}{|\vec{k}|^3},$$

where $\mathscr{P}_0$ is a function of $|\vec{k}|\tau$ only.

Proof. Noting that $A(\tau,\kappa,p)$ is a function of $\kappa\tau$ and $p\tau$ only and performing the ε-limit in (9.9) inside the integral, this can be seen by the substitution $x = p\tau$ in (9.9). □

We would like to improve the estimate of $P_0(\tau,\vec{k})$ for τ close to zero. Adhering to our previous strategy, we shall first give a new estimate for $A(\tau,k,p)$:

Lemma 9.4. *The auxiliary function $A(\tau,\kappa,p)$ is bounded by*

$$\left|A(\tau,\kappa,p)\right| \le -2\frac{\kappa^2\tau}{|p|}, \quad p \ne 0, \quad \tau < 0.$$
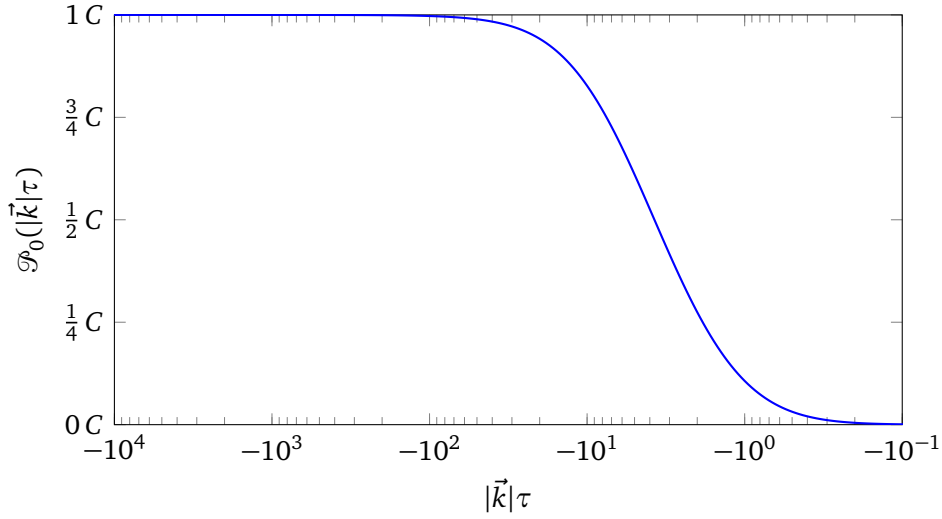
Figure 9.2. Logarithmic plot of the rescaled power spectrum $\mathscr{P}_0(|\vec{k}|\tau)$, where C is the same proportionality constant as in Prop. 9.2.

Proof. Recalling the form of A given in (9.10) and integrating by parts, where we use that $e^{-ip\tau_1} = i p^{-1} \partial_{\tau_1} e^{-ip\tau_1}$, we find

$$A(\tau, \kappa, p) = \frac{i\kappa^2\tau^2}{p} \int_{-\infty}^{\tau} \left(\frac{1}{\tau_1^2} \cos\left(\kappa\left(\tau - \tau_1\right)\right) + \frac{2}{\kappa\tau_1^3} \sin\left(\kappa\left(\tau - \tau_1\right)\right) \right) e^{-ip\tau_1} \, d\tau_1.$$

We then take the absolute value and estimate the trigonometric functions, which gives us a bound on A, namely

$$\left| A(\tau, \kappa, p) \right| \le \frac{\kappa^2\tau^2}{|p|} \int_{-\infty}^{\tau} \left(\frac{1}{\tau_1^2} - 2\frac{\tau - \tau_1}{\tau_1^3} \right) d\tau_1 = -2\frac{\kappa^2\tau}{|p|}. \qquad \square$$

Performing the integration in p analogously to the second part of proposition (9.2), the last lemma immediately leads to a corresponding bound for P_0:

Proposition 9.5. *The leading contribution P_0 of the power spectrum of the potential Φ satisfies the inequality*

$$\left| P_0(\tau, \vec{k}) \right| \le \frac{m^4}{36\pi^2} \frac{\tau^2}{|\vec{k}|}$$

and therefore, in particular, $P_0(0, \vec{k}) = 0$.

The rescaled power spectrum $\mathscr{P}_0(|\vec{k}|\tau)$ can be analyzed numerically and a plot is shown in Fig. 9.2. It clearly exhibits the asymptotic behaviour of P_0 discussed in Props. 9.2 and 9.5. Note that the horizontal axis is logarithmically scaled to highlight the behavior of $\mathscr{P}_0$ for small $|\vec{k}|\tau$, which would be concealed by the fast approach of $\mathscr{P}_0$ to its bound had we used a linear scaling.

In this section we have used the leading singularity[6] of the two-point function of the Bunch–Davis state on a de Sitter universe to compute the influence of quantum

---

[6]Recall that considering only the leading singularity in the Bunch–Davis state also corresponds to the limit of vanishing mass.

matter on the power spectrum of the metric perturbation Φ. We have seen that this results in an almost scale-invariant power spectrum. We stress that such a singularity is not a special feature of the Bunch–Davis state but is common for every Hadamard state. Moreover, although our analysis has been done on a de Sitter universe, similar quantum states have been constructed on universes which are asymptotically de Sitter spaces in the past [61, 62]. All these states tend to the Bunch–Davis state for $\tau \to -\infty$ and are of Hadamard form.

9.2.3   Non-Gaussianities of the metric perturbations

It follows from (9.3c) that the n-point correlation for Φ will, in general, not vanish. Also for odd n they will be different from zero and hence Φ is not a Gaussian random field. As a first measure of the non-Gaussianity of Φ one usually calculates its three-point correlation function or the corresponding bispectrum B:

$$
\langle \Phi(\tau, \vec{x}_1)\, \Phi(\tau, \vec{x}_2)\, \Phi(\tau, \vec{x}_3) \rangle \doteq \frac{1}{(2\pi)^9} \iiint_{\mathbb{R}^9} \delta(\vec{k}_1 + \vec{k}_2 + \vec{k}_3) B(\tau, \vec{k}_1, \vec{k}_2, \vec{k}_3)
$$
$$
\times\, \mathrm{e}^{\mathrm{i}(\vec{k}_1 \cdot \vec{x}_1 + \vec{k}_2 \cdot \vec{x}_2 + \vec{k}_3 \cdot \vec{x}_3)}\, \mathrm{d}\vec{k}_1\, \mathrm{d}\vec{k}_2\, \mathrm{d}\vec{k}_3.
$$

Assuming nonzero $\vec{k}_1$, $\vec{k}_2$ and $\vec{k}_3$, we will derive the form of the bispectrum B considering (as above) only the contribution due to the leading singularity of the Bunch–Davis state, which we will denote by B_0. We will follow the same steps that lead us to the calculation of the power spectrum in the previous section. That is, we apply the retarded propagator Δ_{ret} of (9.5) as in (9.6) to the right-hand side of (9.3c) for $n = 3$ to obtain an equation for Φ and insert for the two-point distribution the conformally rescaled two-point distribution of the massless Minkowski vacuum. The result can again be expressed in terms of the auxiliary function A defined in (9.10):

$$
B_0(\tau, \vec{k}_1, \vec{k}_2, \vec{k}_3) = \lim_{\varepsilon \to 0^+} \frac{m^6}{32\sqrt{3}\vec{k}_1^2 \vec{k}_2^2 \vec{k}_3^2} \int_{\mathbb{R}^3} \left(\frac{\mathrm{e}^{-\varepsilon\left(\omega_{\vec{p}}(-\vec{k}_1) + \omega_{\vec{p}}(\vec{k}_3) + |\vec{p}|\right)}}{\omega_{\vec{p}}(-\vec{k}_1)\omega_{\vec{p}}(\vec{k}_3)|\vec{p}|} \right.
$$
$$
\times A\big(\tau, \kappa_1, \omega_{\vec{p}}(-\vec{k}_1) + |\vec{p}|\big) A\big(\tau, \kappa_3, -\omega_{\vec{p}}(\vec{k}_3) - |\vec{p}|\big) \qquad (9.13)
$$
$$
\left. \times A\big(\tau, \kappa_2, \omega_{\vec{p}}(\vec{k}_3) - \omega_{\vec{p}}(-\vec{k}_1)\big) + \text{permutations} \right) \mathrm{d}\vec{p},
$$

where $\kappa_i \doteq |\vec{k}_i|/\sqrt{3}$, $\omega_{\vec{p}}(\vec{k}) \doteq |\vec{k} + \vec{p}|$ and the sum is over all permutations of $1, 2, 3$.

We can apply the same bound on A which has been used in the previous section to bound the power spectrum P_0 to produce a bound on the integrand of B_0 almost everywhere.[7] Nevertheless, the singularity in the integrand in (9.13) is integrable, *i.e.*, B_0 is bounded. As a consequence we can perform the limit $\varepsilon \to 0^+$ inside the integral.

---

[7]We cannot bound the integrand of B_0 everywhere using (9.11) because $|\vec{k}_2|/\sqrt{3} \neq \left|\omega_{\vec{p}}(\vec{k}_3) - \omega_{\vec{p}}(-\vec{k}_1)\right|$ (and permutations) does not hold everywhere.

Proposition 9.6. *The leading contribution B_0 of the bispectrum of the metric perturbation Φ has the form*

$$B_0(\tau, \vec{k}_1, \vec{k}_2, \vec{k}_3) = \frac{\mathscr{B}_0(k_1\tau, k_2\tau, k_3\tau)}{k_1^2\, k_2^2\, k_3^2},$$

where $\mathscr{B}_0$ is a function of $k_1\tau$, $k_2\tau$, and $k_3\tau$ only and $k_i = |\vec{k}_i|$.

Proof. Analogously to Prop. 9.3, we note that after a change of variables $\vec{x} = \tau\, \vec{p}$ the integrand in (9.13) is a function of $k_1\tau$, $k_2\tau$, and $k_3\tau$ only. $\square$

To finish our discussion about non-Gaussianities, we notice that, although the employed quantum field is a linear one, we obtained a three-point function for Φ which is similar to the one obtained by Maldacena [150] who has quantized metric perturbations outside the linear approximation.

9.3   Outlook

In this chapter the influence of quantum matter fluctuations on metric perturbations over de Sitter backgrounds were analyzed. We used techniques proper of quantum field theory on curved spacetime to regularize the stress-energy tensor and to compute its fluctuations. In particular, we interpreted the perturbations of the curvature tensors as the realization of a stochastic field. We then obtained the n-point distributions of such a stochastic field as induced by the n-point distributions of a quantum stress tensor by means of semiclassical Einstein equations.

We also noticed that, while the expectation value of the stress-energy tensor is characterized by renormalization ambiguities, this is no longer the case when fluctuations are considered. Hence the obtained results are independent on the particular regularization used to define the stress tensor.

In order to keep superficial contact with literature on inflation, we investigated perturbations of the scalar curvature generated by a Newtonian metric perturbation, which is related to the standard Bardeen potentials. However, the considered model is certainly oversimplified to cover any real situation and is not gauge invariant.

Within this model it was possible to recover an almost-Harrison–Zel'dovich power spectrum for the considered metric perturbation. Furthermore, the amplitude of such a power spectrum depends on the field mass which is a free parameter in our model and can be fixed independently of H. At the same time, since it does not depend on the Hubble parameter of the background metric, this indicates that it is not a special feature of de Sitter space. At least close to the initial singularity, the obtained result depends only on the form of the most singular part of the two-point function of the considered Bunch–Davis state. We thus argue that a similar feature is present in every Hadamard state and for backgrounds which are only asymptotically de Sitter in the past.

Finally we notice that, since the stress-energy tensor is not linear in the field, its probability distribution cannot be of Gaussian nature. Thus we showed that non-Gaussianities arise naturally in this picture.

Conclusions

In this thesis aspects of the backreaction of quantum matter fields on the curvature of spacetime were discussed. The main results in this direction were discussed in Chaps. 8 and 9: the existence of local and global solutions of the semiclassical Einstein equation on cosmological spacetimes, and the a coupling the fluctuations of the quantum matter field to a Newtonianly perturbed de Sitter spacetime. Further results presented are the enumerative combinatorics of the run structure of permutations in Chap. 4 with applications to the moment problem of the Wick square and the stress-energy on Minkowski spacetime.

In each case the problems were not treated in all possible generality, mainly due to the difficulty of constructing Hadamard states on general globally hyperbolic spacetimes but also due to other factors. Nevertheless, we studied the effects of quantum fields on cosmological spacetimes not only because of their relative simplicity but also because of the relevance in cosmology. Therefore the first avenue is not always the generalization of results to more general spacetimes, but also the better understanding of possible effects on this restricted class of spacetimes. For example, we already mentioned in Sect. 8.5 that the results are restricted to the conformally coupled scalar field with a certain choice for the renormalization freedom as other choices can lead to equations involving higher than second-order derivatives of the metric and ask for slightly different approach. However, it would be desirable to understand this problem also for non-conformal coupling and discuss the full dependence of the solutions to the Einstein equation on the renormalization freedom.

In the case of results on the metric fluctuations induced by quantum matter fluctuations as presented in Chap. 9, we were even more restrictive and the discussion is mostly based on the special case of a Newtonianly perturbed de Sitter spacetime. While straightforward generalization of this idea to asymptotically de Sitter spacetimes are possible and were already published in [60], the next step should be to gain a clearer physical and mathematical motivation of the used equations. A development in this direction is [75], but also this work should only be seen as a first step. In any case, as soon as one attempts to take into account the fluctuations of the stress-energy tensor one is faced with the limitations of the semiclassical Einstein equation and any attempts to generalize them, even if well-motivated, remains speculative in the absence of an accepted theory of quantum gravity.

On one hand, when attempting to study quantum field theory in a mathematically rigorous fashion one sees even clearer the non-uniqueness of many constructions and one is confronted with many choices: Are all Hadamard states physically sensible? What topology should be chosen for the algebra of quantum fields? What is the appropriate gauge freedom for the electromagnetic potential on non-contractible spacetimes? Is it reasonable to work with an algebra of unbouded operators such as the field algebra or should one always use a C^*-algebra? There are many more questions of this kind and they require further mathematical and physical investigations but also intuition. Quantum field theory and the quest for a quantum gravity is and

will continue to be not only a research effort and playground of physicists but also one of mathematicians.

On the other hand, many aspects of quantum field theory are now conceptually and mathematically very well understood but only a few models have been studied in all their detail. In particular interacting quantum fields on curved spacetimes have received relatively little attention given that already free fields are a complicated matter. Investigations of physically interesting interacting models, using perturbative techniques, are largely absent from the literature and deserve more attention.

For these reasons one can expect that the field of quantum field theory (on curved spacetimes) will remain an interesting field of study for many more years to come.

Acknowledgements

First of all, I would like to thank the University of Genova for providing me with the opportunity to carry out this work. I am especially indebted to my amazing supervisor Nicola Pinamonti, whom I could always seek for advice and guidance, also on fields unrelated to science; only the collaboration with him made much of this thesis possible.

Second, my thanks go to the University of York for their hospitality during a 6 months stay (September 2013 – March 2014) to collaborate with Chris Fewster. In many enlightening discussions with him I could broaden my horizon further and complete parts of this work.

Third, I would like to thank Valter Moretti for agreeing to be my examiner and thereby taking upon him the task to read this long thesis.

Furthermore, I would like to thank my friends and colleagues from Genova, York and the Local Quantum Physics community. Not only am I very grateful for their friendship, unforgettable shared memories and good times far from home, but also countless inspiring conversations on science and other topics.

Last, but not least, my gratitude also goes to my family, especially my parents, who always supported me, even if the do not understand what I am doing.

Bibliography

Works by the author

[1] Dappiaggi, C. and Siemssen, D. "Hadamard states for the vector potential on asymptotically flat spacetimes". *Reviews in Mathematical Physics* 25.1, 1350002 (2013). arXiv:1106.5575 [gr-qc].

[2] Fewster, C. J. and Schenkel, A. "Locally covariant quantum field theory with external sources". *Annales Henri Poincaré* (2014). arXiv:1402.2436 [math-ph].

[3] Pinamonti, N. and Siemssen, D. "Global existence of solutions of the semiclassical Einstein equation for cosmological spacetimes". *Communications in Mathematical Physics* 334.1 (2015), pp. 171–191. arXiv:1309.6303 [math-ph].

[4] Pinamonti, N. and Siemssen, D. "Scale-invariant curvature fluctuations from an extended semiclassical gravity". *Journal of Mathematical Physics* 56, 022303 (2015). arXiv:1303.3241 [gr-qc].

[5] Siemssen, D. "Quantization of the electromagnetic potential in asymptotically flat spacetimes". Diplomarbeit. Universität Hamburg, 2011.

Works by other authors

[6] Abraham, R., Marsden, J. E., and Ratiu, T. S. *Manifolds, Tensor Analysis, and Applications*. Applied Mathematical Sciences 75. Springer, 1988.

[7] Agulló, I. et al. "Reexamination of the power spectrum in de Sitter inflation". *Physical Review Letters* 101.17, 171301 (2008). arXiv:0806.0034 [gr-qc].

[8] Agulló, I. et al. "Remarks on the renormalization of primordial cosmological perturbations". *Physical Review D* 84.10, 107304 (2011). arXiv:1108.0949 [gr-qc].

[9] Agulló, I. et al. "Revising the predictions of inflation for the cosmic microwave background anisotropies". *Physical Review Letters* 103, 061301 (2009). arXiv:0901.0439 [astro-ph.CO].

[10] Allen, B. "Vacuum states in de Sitter space". *Physical Review D* 32.12 (1985), pp. 3136–3149.

[11] Allen, B. and Folacci, A. "Massless minimally coupled scalar field in de Sitter space". *Physical Review D* 35.12 (1987), pp. 3771–3778.

[12] Anderson, P. R. "Effects of quantum fields on singularities and particle horizons in the early universe". *Physical Review D* 28.2 (1983), pp. 271–285.

[13] Anderson, P. R. "Effects of quantum fields on singularities and particle horizons in the early universe. II". *Physical Review D* 29.4 (1984), pp. 615–627.

[14]   Anderson, P. R. "Effects of quantum fields on singularities and particle hori-
       zons in the early universe. III. The conformally coupled massive scalar field".
       Physical Review D 32.6 (1985), pp. 1302–1315.

[15]   Anderson, P. R. "Effects of quantum fields on singularities and particle horizons
       in the early universe. IV. Initially empty universes". *Physical Review D* 33.6
       (1986), pp. 1567–1575.

[16]   André, D. "Mémoire sur les séquences des permutations circulaires". *Bulletin
       de la Société Mathématique de France* 23 (1895), pp. 122–184. eprint: http:
       //eudml.org/doc/85822.

[17]   André, D. "Sur les permutations alternées". *Journal de Mathématiques Pures
       et Appliquées* 7 (1881), pp. 167–184. eprint: https://eudml.org/doc/
       233984.

[18]   Aoyama, T. et al. "Tenth-order QED contribution to the electron g-2 and an
       improved value of the fine structure constant". *Physical Review Letters* 109.11,
       111807 (2012). arXiv:1205.5368 [hep-ph].

[19]   Ashtekar, A., Pretorius, F., and Ramazanoğlu, F. M. "Surprises in the evap-
       oration of 2D black holes". *Physical Review Letters* 106.16, 161303 (2011).
       arXiv:1011.6442 [gr-qc].

[20]   Averbukh, V. I. and Smolyanov, O. G. "The various definitions of the derivative
       in linear topological spaces". *Russian Mathematical Surveys* 23.4 (1968), p. 67.

[21]   Avramidi, I. G. *Heat Kernel and Quantum Gravity*. Lecture Notes in Physics 64.
       Springer, 2000.

[22]   Avramidi, I. G. "Covariant methods for the calculation of the effective ac-
       tion in quantum field theory and investigation of higher derivative quantum
       gravity". PhD thesis. Moscow State Lomonosov University, 1986. arXiv:hep-
       th/9510140 [hep-th].

[23]   Bär, C. and Ginoux, N. "Classical and quantum fields on lorentzian manifolds".
       In: *Global Differential Geometry*. Ed. by Bär, C., Lohkamp, J., and Schwarz, M.
       Vol. 17.2. Springer Proceedings in Mathematics. Springer, 2012, pp. 359–400.
       arXiv:1104.1158 [math-ph].

[24]   Bär, C., Ginoux, N., and Pfäffle, F. *Wave Equations on Lorentzian Manifolds
       and Quantization*. ESI Lectures in Mathematical Physics. Zürich: European
       Mathematical Society, 2007. arXiv:0806.1036 [math.DG].

[25]   Bardeen, J. M. "Gauge invariant cosmological perturbations". *Physical Review
       D* 22.8 (1980), pp. 1882–1905.

[26]   Bartolo, N. et al. "Non-Gaussianity from inflation: theory and observations".
       Physics Reports 402 (2004), pp. 103–266. arXiv:astro-ph/0406398.

[27]   Beem, J. K., Ehrlich, P. E., and Easley, K. L. *Global Lorentzian Geometry*. 2nd ed.
       Pure and Applied Mathematics. 202: Marcel Dekker, 1996.

[28]   Benini, M. "Relative Cauchy evolution for spin 1 fields". Master Thesis. Univer-
       sità Pavia, 2011. arXiv:1111.6471 [math-ph].

[29] Benini, M., Dappiaggi, C., and Murro, S. "Radiative observables for linearized gravity on asymptotically flat spacetimes and their boundary induced states". *Journal of Mathematical Physics* 55.8, 082301 (2014). arXiv:1404.4551 [gr-qc].

[30] Benini, M., Dappiaggi, C., and Schenkel, A. "Quantized Abelian principal connections on Lorentzian manifolds". *Communications in Mathematical Physics* 330 (2014), pp. 123–152. arXiv:1303.2515 [math-ph].

[31] Benini, M., Dappiaggi, C., and Schenkel, A. "Quantum field theory on affine bundles". *Annales Henri Poincaré* 15.1 (2014), pp. 171–211. arXiv:1210.3457 [math-ph].

[32] Benini, M. et al. "A C*-algebra for quantized principal U(1)-connections on globally hyperbolic Lorentzian manifolds". *Communications in Mathematical Physics* (2014). arXiv:1307.3052 [math-ph].

[33] Bennett, C. L. et al. "Nine-year Wilkinson Microwave Anisotropy Probe (WMAP) observations: final maps and results". *The Astrophysical Journal Supplement Series* 208.2, 20 (2012). arXiv:1212.5225 [astro-ph.CO].

[34] Bernal, A. N. and Sánchez, M. "On smooth Cauchy hypersurfaces and Geroch's splitting theorem". *Communications in Mathematical Physics* 243.3 (2003), pp. 461–470. arXiv:gr-qc/0306108.

[35] Bernal, A. N. and Sánchez, M. "Smoothness of time functions and the metric splitting of globally hyperbolic spacetimes". *Communications in Mathematical Physics* 257.1 (2005), pp. 43–50. arXiv:gr-qc/0401112.

[36] Birrell, N. D. and Davies, P. C. W. *Quantum Fields in Curved Space*. Cambridge Monographs on Mathematical Physics. Cambridge University Press, 1984.

[37] Bóna, M. *Combinatorics of Permutations*. 2nd ed. Discrete Mathematics and Its Applications. Taylor & Francis, 2012.

[38] Borchers, H.-J. "Algebraic aspects of Wightman field theory". In: *Statistical Mechanics and Field Theory*. Ed. by Sen, R. N. and Weil, C. John Wiley & Sons, 1972, pp. 31–89.

[39] Borchers, H.-J. "On the structure of the algebra of field operators". *Il Nuovo Cimento* 24.2 (1962), pp. 214–236.

[40] Brewin, L. "Riemann normal coordinate expansions using Cadabra". *Classical and Quantum Gravity* 26.17, 175017 (2009). arXiv:0903.2087 [gr-qc].

[41] Brouder, C., Viet Dang, N., and Hélein, F. "A smooth introduction to the wavefront set" (2014). arXiv:1404.1778 [math-ph].

[42] Brown, F. C. S., Fink, T. M. A., and Willbrand, K. "On arithmetic and asymptotic properties of up–down numbers". *Discrete Mathematics* 307.14 (2007), pp. 1722–1736. arXiv:math/0607763 [math.CO].

[43] Brunetti, R., Dütsch, M., and Fredenhagen, K. "Perturbative algebraic quantum field theory and the renormalization groups". *Advances in Theoretical and Mathematical Physics* 13.5 (2009), pp. 1541–1599. arXiv:0901.2038 [math-ph].

[44] Brunetti, R. and Fredenhagen, K. "Microlocal analysis and interacting quantum field theories: Renormalization on physical backgrounds". *Communications in Mathematical Physics* 208.3 (2000), pp. 623–661. arXiv:math-ph/9903028.

[45] Brunetti, R., Fredenhagen, K., and Köhler, M. "The microlocal spectrum condition and Wick polynomials of free fields on curved spacetimes". *Communications in Mathematical Physics* 180.3 (1996), pp. 633–652.

[46] Brunetti, R., Fredenhagen, K., and Verch, R. "The generally covariant locality principle – a new paradigm for local quantum field theory". *Communications in Mathematical Physics* 237.1 (2003), pp. 31–68. arXiv:math-ph/0112041.

[47] Brunetti, R. and Ruzzi, G. "Superselection sectors and general covariance. i". *Communications in Mathematical Physics* 270.1 (2007), pp. 69–108. arXiv:gr-qc/0511118.

[48] Brunetti, R. et al. "The locality axiom in quantum field theory and tensor products of c^*-algebras". *Reviews in Mathematical Physics* 26, 1450010 (2014). arXiv:1206.5484 [math-ph].

[49] Buchholz, D. and Schlemmer, J. "Local temperature in curved spacetime". *Classical and Quantum Gravity* 24.7, F25 (2007). arXiv:arXiv:gr-qc/0608133.

[50] Bunch, T. S. "Adiabatic regularisation for scalar fields with arbitrary coupling to the scalar curvature". *Journal of Physics A: Mathematical and General* 13.4 (1980), pp. 1297–1310.

[51] Bunch, T. S. and Davies, P. C. W. "Quantum field theory in de Sitter space: renormalization by point-splitting". *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences* 360.1700 (1978), pp. 117–134.

[52] Busch, P. "Unsharp localization and causality in relativistic quantum theory". *Journal of Physics A: Mathematical and General* 32.37 (1999), pp. 6535–6546. arXiv:1303.6649 [quant-ph].

[53] Cantor, M. "Elliptic operators and the decomposition of tensor fields". *Bulletin of the American Mathematical Society* 5.3 (1981), pp. 235–262.

[54] Dabrowski, Y. and Brouder, C. "Functional properties of Hörmander's space of distributions having a specified wavefront set". *Communications in Mathematical Physics* 332.3 (2014), pp. 1345–1380. arXiv:1308.1061 [math-ph].

[55] Dappiaggi, C. "Remarks on the Reeh–Schlieder property for higher spin free fields on curved spacetimes". *Reviews in Mathematical Physics* 23.10 (2011), pp. 1035–1062. arXiv:1102.5270 [math-ph].

[56] Dappiaggi, C., Fredenhagen, K., and Pinamonti, N. "Stable cosmological models driven by a free quantum scalar field". *Physical Review D* 77.10, 104015 (2008). arXiv:0801.2850 [gr-qc].

[57] Dappiaggi, C., Hack, T.-P., and Pinamonti, N. "Approximate KMS states for scalar and spinor fields in Friedmann–Robertson–Walker spacetimes". *Annales Henri Poincaré* 12.8 (2011), pp. 1449–1489. arXiv:1009.5179 [gr-qc].

[58] Dappiaggi, C., Hack, T.-P., and Pinamonti, N. "The extended algebra of observables for Dirac fields and the trace anomaly of their stress-energy tensor". *Reviews in Mathematical Physics* 21.10 (2009), pp. 1241–1312. arXiv:0904.0612 [math-ph].

[59] Dappiaggi, C. and Lang, B. "Quantization of Maxwell's equations on curved backgrounds and general local covariance". *Letters in Mathematical Physics* 101.3 (2012), pp. 265–287. arXiv:1104.1374 [gr-qc].

[60] Dappiaggi, C. and Melati, A. "Curvature fluctuations on asymptotically de Sitter spacetimes via the semiclassical Einstein's equations". *Classical and Quantum Gravity* 31.23, 235006 (2014). arXiv:1406.2223 [gr-qc].

[61] Dappiaggi, C., Moretti, V., and Pinamonti, N. "Cosmological horizons and reconstruction of quantum field theories". *Communications in Mathematical Physics* 285.3 (2009), pp. 1129–1163. arXiv:0712.1770 [gr-qc].

[62] Dappiaggi, C., Moretti, V., and Pinamonti, N. "Distinguished quantum states in a class of cosmological spacetimes and their Hadamard property". *Journal of Mathematical Physics* 50, 062304 (2009). arXiv:0812.4033 [gr-qc].

[63] Dappiaggi, C., Moretti, V., and Pinamonti, N. "Rigorous construction and Hadamard property of the Unruh state in Schwarzschild spacetime". *Advances in Theoretical and Mathematical Physics* 15.2 (2011), pp. 1–93. arXiv:0907.1034 [gr-qc].

[64] Dappiaggi, C., Moretti, V., and Pinamonti, N. "Rigorous steps towards holography in asymptotically flat spacetimes". *Reviews in Mathematical Physics* 18.4 (2006), pp. 349–415. arXiv:gr-qc/0506069.

[65] Dappiaggi, C., Pinamonti, N., and Porrmann, M. "Local causal structures, Hadamard states and the principle of local covariance in quantum field theory". *Communications in Mathematical Physics* 304.2 (2011), pp. 459–498. arXiv:1001.0858 [hep-th].

[66] Dappiaggi, C. et al. "Dark energy from quantum matter" (2010). arXiv:1007.5009 [astro-ph.CO].

[67] Davies, P. C. W. "Scalar production in Schwarzschild and Rindler metrics". *Journal of Physics A: Mathematical and General* 8 (1975), pp. 609–616.

[68] Davies, P. C. W. et al. "Energy-momentum tensor of a massless scalar quantum field in a Robertson–Walker universe". *Annals of Physics* 109.1 (1977), pp. 108–142.

[69] Décanini, Y. and Folacci, A. "Off-diagonal coefficients of the DeWitt–Schwinger and Hadamard representations of the Feynman propagator". *Physical Review D* 73.4, 044027 (2006). arXiv:gr-qc/0511115.

[70] Degner, A. "Properties of states of low energy on cosmological spacetimes". PhD thesis. Universität Hamburg, 2013.

[71] Degner, A. and Verch, R. "Cosmological particle creation in states of low energy". *Journal of Mathematical Physics* 51.2, 022302 (2010). arXiv:0904.1273 [gr-qc].

[72] Dimock, J. "Algebras of local observables on a manifold". *Communications in Mathematical Physics* 77.3 (1980), pp. 219–228.

[73] Dimock, J. "Dirac quantum fields on a manifold". *Transactions of the American Mathematical Society* 269.133–147 (1982).

[74] Dimock, J. "Quantized electromagnetic field on a manifold". *Reviews in Mathematical Physics* 4.2 (1992), pp. 223–233.

[75] Drago, N. and Pinamonti, N. "Influence of quantum matter fluctuations on geodesic deviation". *Journal of Physics A: Mathematical and General* 47.37, 375202 (2014). arXiv:1402.4265 [math-ph].

[76] Durrer, R. *The Cosmic Microwave Background*. Cambridge University Press, 2008.

[77] Durrer, R., Marozzi, G., and Rinaldi, M. "On adiabatic renormalization of inflationary perturbations". *Physical Review D* 80.6, 065024 (2009). arXiv:0906.4772 [astro-ph.CO].

[78] Ehlers, J., Geren, P., and Sachs, R. K. "Isotropic solutions of the Einstein–Liouville equations". *Journal of Mathematical Physics* 9.9 (1968), pp. 1344–1349.

[79] Ellis, G. F. R. and Bruni, M. "Covariant and gauge-invariant approach to cosmological density fluctuations". *Physical Review D* 40.6 (1989), pp. 1804–1818.

[80] Ellis, G. F. R., Maartens, R., and MacCallum, M. A. H. *Relativistic Cosmology*. Cambridge University Press, 2012.

[81] Eltzner, B. "Quantization of perturbations in inflation" (2013). arXiv:1302.5358 [gr-qc].

[82] Eltzner, B. "Thermal equilibrium on curved spacetimes and linear cosmological perturbation theory". PhD thesis. Universität Leipzig, 2013.

[83] Eltzner, B. and Gottschalk, H. "Dynamical backreaction in Robertson–Walker spacetime". *Reviews in Mathematical Physics* 23.5 (2011), pp. 531–551. arXiv:1003.3630 [math-ph].

[84] Engquist, B. et al., eds. *Highly Oscillatory Problems*. London Mathematical Society Lecture Note Series 366. Cambridge University Press, 2009.

[85] Fewster, C. J. in preperation.

[86] Fewster, C. J. "Endomorphisms and automorphisms of locally covariant quantum field theories". *Reviews in Mathematical Physics* 25, 1350008 (2013). arXiv:1201.3295 [math-ph].

[87] Fewster, C. J. "Energy inequalities in quantum field theory". In: *XIVth International Congress on Mathematical Physics*. Ed. by Zambrini, J.-C. 2006, pp. 559–568. arXiv:math-ph/0501073.

[88] Fewster, C. J. *Lectures on Quantum Energy Inequalities*. 2012. arXiv:1208.5399v1 [gr-qc].

[89] Fewster, C. J. "On the notion of 'the same physics in all spacetimes'". In: *Quantum Field Theory and Gravity. Conceptual and Mathematical Advances in the Search for a Unified Framework*. Ed. by Finster, F. et al. Birkhäuser, 2012, pp. 207–227. arXiv:1105.6202 [math-ph].

[90] Fewster, C. J., Ford, L. H., and Roman, T. A. "Probability distributions for quantum stress tensors in four dimensions". *Physical Review D* 85.12, 125038 (2012). arXiv:1204.3570 [quant-ph].

[91] Fewster, C. J., Ford, L. H., and Roman, T. A. "Probability distributions of smeared quantum stress tensors". *Physical Review D* 81.12, 121901 (2010). arXiv:1004.0179 [quant-ph].

[92] Fewster, C. J. and Hunt, D. S. "Quantization of linearized gravity in cosmological vacuum spacetimes". *Reviews in Mathematical Physics* 25.2, 1330003 (2013). arXiv:1203.0261 [math-ph].

[93] Fewster, C. J. and Lang, B. "Dynamical locality of the free Maxwell field" (2014). arXiv:1403.7083 [math-ph].

[94] Fewster, C. J. and Pfenning, M. J. "A quantum weak energy inequality for spin-one fields in curved space–time". *Journal of Mathematical Physics* 44.10 (2003), pp. 4480–4513. arXiv:gr-qc/0303106.

[95] Fewster, C. J. and Siemssen, D. "Enumerating permutations by their run structure". *The Electronic Journal of Combinatorics* 21.4, P4.18 (2014). arXiv:1403.1723 [math.CO].

[96] Fewster, C. J. and Smith, C. J. "Absolute quantum energy inequalities in curved spacetime". *Annales Henri Poincaré* 9.3 (2008), pp. 425–455. arXiv:gr-qc/0702056.

[97] Fewster, C. J. and Verch, R. "Dynamical locality and covariance: what makes a physical theory the same in all spacetimes?" *Annales Henri Poincaré* 13.7 (2012), pp. 1613–1674. arXiv:1106.4785 [math-ph].

[98] Fewster, C. J. and Verch, R. "Dynamical locality of the free scalar field". *Annales Henri Poincaré* 13 (2012), pp. 1675–1709. arXiv:1109.6732 [math-ph].

[99] Fragoulopoulou, M. *Topological Algebras with Involution*. Mathematics Studies 200. North-Holland, 2005.

[100] Fredenhagen, K. and Haag, R. "On the derivation of Hawking radiation associated with the formation of a black hole". *Communications in Mathematical Physics* 127.2 (1990), pp. 273–284.

[101] Fredenhagen, K. and Rejzner, K. "Batalin–Vilkovisky formalism in perturbative algebraic quantum field theory". *Communications in Mathematical Physics* 317 (2013), pp. 697–725. arXiv:1110.5232 [math-ph].

[102] Fredenhagen, K. and Rejzner, K. *Perturbative Algebraic Quantum Field Theory*. 2012. arXiv:1208.1428v1 [math-ph].

[103] Friedlander, F. G. *The Wave Equation on a Curved Space-Time*. Cambridge University Press, 1975.

[104] Fulling, S. A. *Aspects of Quantum Field Theory in Curved Space-Time*. London Mathematical Society Student Texts 17. Cambridge University Press, 1989.

[105] Fulling, S. A. "Nonuniqueness of canonical field quantization in Riemannian space-time". *Physical Review D* 7.10 (1973), pp. 2850–2862.

[106] Gérard, C. and Wrochna, M. "Construction of Hadamard states by characteristic Cauchy problem" (2014). arXiv:1409.6691 [math-ph].

[107] Gradshteyn, I. S. and Ryzhin, I. M. *Table of Integrals, Series and Products*. Ed. by Jeffrey, A. and Zwillinger, D. Vol. 7. Academic Press, 2007.

[108] Guth, A. H. and Pi, S.-Y. "Fluctuations in the new inflationary universe". *Physical Review Letters* 49.15 (1982), pp. 1110–1113.

[109] Haag, R. and Kastler, D. "An algebraic approach to quantum field theory". *Journal of Mathematical Physics* 5.7 (1964), pp. 848–861.

[110] Haag, R. *Local Quantum Physics. Fields, Particles, Algebras*. 2nd ed. Springer, 1996.

[111] Haag, R., Narnhofer, H., and Stein, U. "On quantum field theory in gravitational background". *Communications in Mathematical Physics* 94.2 (1984), pp. 219–238.

[112] Hack, T.-P. "On the backreaction of scalar and spinor quantum fields in curved spacetimes". PhD thesis. Universität Hamburg, 2010. arXiv:1008.1776 [gr-qc].

[113] Hack, T.-P. "Quantization of the linearised Einstein–Klein–Gordon system on arbitrary backgrounds and the special case of perturbations in inflation" (2014). arXiv:1403.3957 [gr-qc].

[114] Hack, T.-P. "The ΛCDM-model in quantum field theory on curved spacetime and dark radiation" (2013). arXiv:1306.3074 [gr-qc].

[115] Hack, T.-P. and Schenkel, A. "Linear bosonic and fermionic quantum gauge theories on curved spacetimes". *General Relativity and Gravitation* 45.5 (2013), pp. 877–910. arXiv:1205.3484 [math-ph].

[116] Hamilton, R. S. "The inverse function theorem of Nash and Moser". *Bulletin of the American Mathematical Society* 7.1 (1982), pp. 65–222.

[117] Hanneke, D., Fogwell, S., and Gabrielse, G. "New measurement of the electron magnetic moment and the fine structure constant". *Physical Review Letters* 100.12, 120801 (2008). arXiv:0801.1134 [physics.atom-ph].

[118] Hawking, S. W. "Particle creation by black holes". *Communications in Mathematical Physics* 43.3 (1975), pp. 199–220.

[119] Hawking, S. W. "Perturbations of an expanding universe". *Astrophysical Journal* 145 (1966), pp. 544–554.

[120] Hawking, S. W. "The development of irregularities in a single bubble inflationary universe". *Physics Letters B* 115.4 (1982), pp. 295–297.

[121]   Hinshaw, G., Larson, D., et al. "Nine-year Wilkinson Microwave Anisotropy Probe (WMAP) observations: cosmological parameter results". *The Astrophysical Journal Supplement Series* 208.2, 19 (2012). arXiv:1212.5226 [astro-ph.CO].

[122]   Hollands, S. and Wald, R. M. "Conservation of the stress tensor in perturbative interacting quantum field theory in curved spacetimes". *Reviews in Mathematical Physics* 17.3 (2005), pp. 277–311. arXiv:gr-qc/0404074.

[123]   Hollands, S. and Wald, R. M. "Existence of local covariant time ordered products of quantum fields in curved spacetime". *Communications in Mathematical Physics* 231.2 (2002), pp. 309–345. arXiv:gr-qc/0111108.

[124]   Hollands, S. and Wald, R. M. "Local Wick polynomials and time ordered products of quantum fields in curved spacetime". *Communications in Mathematical Physics* 223.2 (2001), pp. 289–326. arXiv:gr-qc/0103074.

[125]   Hörmander, L. *The Analysis of Linear Partial Differential Operators*. Vol. 2: *Differential Operators with Constant Coefficients*. Grundlehren der mathematischen Wissenschaften 257. Springer, 1983.

[126]   Hörmander, L. *The Analysis of Linear Partial Differential Operators*. Vol. 1: *Distribution Theory and Fourier Analysis*. 2nd ed. Grundlehren der mathematischen Wissenschaften 256. Springer, 1990.

[127]   Hörmander, L. *The Analysis of Linear Partial Differential Operators*. Vol. 4: *Fourier Integral Operators*. Grundlehren der mathematischen Wissenschaften 275. Springer, 1985.

[128]   Hörmander, L. *The Analysis of Linear Partial Differential Operators*. Vol. 3: *Pseudo-Differential Operators*. Grundlehren der mathematischen Wissenschaften 274. Springer, 1985.

[129]   Hu, B. L. and Verdaguer, E. "Stochastic gravity: theory and applications". *Living Reviews in Relativity* 11.3 (2008).

[130]   Inoue, A. *Tomita–Takesaki Theory in Algebras of Unbounded Operators*. Lecture Notes in Mathematics 1699. Springer, 1998.

[131]   Iserles, A. "Think globally, act locally: solving highly-oscillatory differential equations". *Applied Numerical Mathematics* 43.1–2 (2002), pp. 145–160.

[132]   Iserles, A., Nørsett, S. P., and Olver, S. "Highly oscillatory quadratures: the story so far". In: *Numerical Mathematics and Advanced Applications*. 2006, pp. 97–118.

[133]   Jarchow, H. *Locally Convex Spaces*. Mathematische Leitfäden. B. G. Teubner, 1981.

[134]   Junker, W. and Schrohe, E. "Adiabatic vacuum states on general spacetime manifolds: definition, construction, and physical properties, construction, and physical properties". *Annales Henri Poincaré* 3.6 (2002), pp. 1113–1181. arXiv:math-ph/0109010.

[135]   Kay, B. S. "Quantum fields in curved spacetime: non global hyperbolicity and locality". In: *Operator Algebras and Quantum Field Theory*. Ed. by Doplicher, S. et al. International Press, 1997, pp. 578–588. eprint: gr-qc/9704075.

[136] Kay, B. S. and Wald, R. M. "Theorems on the uniqueness and thermal properties of stationary, nonsingular, quasifree states on spacetimes with a bifurcate Killing horizon". *Physics Reports* 207.2 (1991), pp. 49–136.

[137] Kelley, J. L. *General Topology*. Graduate Texts in Mathematics 27. Springer, 1955.

[138] Kitaev, S. "Introduction to partially ordered patterns". *Discrete Applied Mathematics* 155.8 (2007), pp. 929–944. arXiv:math/0603122 [math.CO].

[139] Knuth, D. E. *Sorting and Searching*. 2nd ed. Vol. 3. The Art of Computer Programming. Addison-Wesley, 1998.

[140] Kodama, H. and Sasaki, M. "Cosmological perturbation theory". *Progress of Theoretical Physics Supplement* 78 (1984), pp. 1–166.

[141] Kofman, L. A., Linde, A. D., and Starobinsky, A. A. "Inflationary universe generated by the combined action of a scalar field and gravitational vacuum polarization". *Physics Letters B* 157.5–6 (1985), pp. 361–367.

[142] Lee, J. M. *Introduction to Smooth Manifolds*. 2nd ed. Graduate Texts in Mathematics 218. Springer, 2013.

[143] Levin, D. "Analysis of a collocation method for integrating rapidly oscillatory functions". *Journal of Computational and Applied Mathematics* 78.1 (1997), pp. 131–138.

[144] Levin, D. "Fast integration of rapidly oscillatory functions". *Journal of Computational and Applied Mathematics* 67.1 (1996), pp. 95–101.

[145] Lichnerowicz, A. "Propagateurs et commutateurs en relativité générale". *Publications Mathématiques de l'I.H.É.S.* 10.1 (1961), pp. 5–56.

[146] Lüders, C. and Roberts, J. E. "Local quasiequivalence and adiabatic vacuum states". *Communications in Mathematical Physics* 134.1 (1990), pp. 29–63.

[147] Lyth, D. H. and Mukherjee, M. "Fluid-flow description of density irregularities in the Universe". *Physical Review D* 38.2 (1988), pp. 485–489.

[148] Maartens, R., Ellis, G. F. R., and Stoeger, W. R. "Limits on anisotropy and inhomogeneity from the cosmic background radiation". *Physical Review D* 51.4 (1995), pp. 1525–1535.

[149] Malament, D. B. "In defense of dogma: Why there cannot be a relativistic quantum mechanics of (localizable) particles". In: *Perspectives on Quantum Reality*. Ed. by Clifton, R. Vol. 57. The University of Western Ontario Series in Philosophy of Science. Springer, 1996, pp. 1–10.

[150] Maldacena, J. "Non-Gaussian features of primordial fluctuations in single field inflationary models". *Journal of High Energy Physics* 5 (2003). arXiv:astro-ph/0210603.

[151] Manuceau, J. et al. "The smallest c^*-algebra for canonical commutations relations". *Communications in Mathematical Physics* 32.3 (1973), pp. 231–243.

[152] Maplesoft, a division of Waterloo Maple Inc. *Maple 16*.

[153] Marozzi, G., Rinaldi, M., and Durrer, R. "On infrared and ultraviolet divergences of cosmological perturbations". *Physical Review D* 83.10, 105017 (2011). arXiv:1102.2206 [astro-ph.CO].

[154] Moretti, V. "Comments on the stress-energy tensor operator in curved spacetime". *Communications in Mathematical Physics* 232.2 (2003), pp. 189–221. arXiv:gr-qc/0109048.

[155] Moretti, V. "Proof of the symmetry of the off-diagonal Hadamard/Seeley–DeWitt's coefficients in C^∞ Lorentzian manifolds by a 'local Wick rotation'". *Communications in Mathematical Physics* 212.1 (2000), pp. 165–189. arXiv:gr-qc/9908068.

[156] Moretti, V. "Proof of the symmetry of the off-diagonal heat-kernel and Hadamard's expansion coefficients in general C^∞ Riemannian manifolds". *Communications in Mathematical Physics* 208.2 (1999), pp. 283–308. arXiv:gr-qc/9902034.

[157] Moretti, V. "Quantum out-states holographically induced by asymptotic flatness: invariance under spacetime symmetries, energy positivity and Hadamard property". *Communications in Mathematical Physics* 279.1 (2008), pp. 31–75. arXiv:gr-qc/0610143.

[158] Moretti, V. *Spectral Theory and Quantum Mechanics. With an Introduction to the Algebraic Formulation*. UNITEXT – La Matematica per il 3 + 2 64. Springer, 2013.

[159] Moretti, V. "Uniqueness Theorem for BMS-Invariant States of Scalar QFT on the Null Boundary of Asymptotically Flat Spacetimes and Bulk-Boundary Observable Algebra Correspondence". *Communications in Mathematical Physics* 268.3 (2006), pp. 727–756. arXiv:gr-qc/0512049.

[160] Mühlhoff, R. "Cauchy problem and Green's functions for first order differential operators and algebraic quantization". *Journal of Mathematical Physics* 52.2, 022303 (2011). arXiv:1001.4091 [math-ph].

[161] Mukhanov, V. F. and Chibisov, G. V. "Quantum fluctuations and a nonsingular universe". *JETP Letters* 33 (1981), pp. 532–535.

[162] Mukhanov, V. F., Feldman, H. A., and Brandenberger, R. H. "Theory of cosmological perturbations". *Physics Reports* 215.5–6 (1992), pp. 203–333.

[163] Neeb, K.-H. *Monastir Summer School: Infinite-Dimensional Lie Groups*. Lecture Notes. 2005.

[164] OEIS Foundation Inc. *The On-Line Encyclopedia of Integer Sequences*. 2013.

[165] Olbermann, H. "States of low energy on Robertson–Walker spacetimes". *Classical and Quantum Gravity* 24.20 (2007), pp. 5011–5030. arXiv:0704.2986 [gr-qc].

[166] Olver, F. W. J. et al., eds. *NIST Handbook of Mathematical Functions*. Cambridge University Press, 2010.

[167] O'Neill, B. *Semi-Riemannian Geometry. With Applications to Relativity*. Pure and Applied Mathematics 103. Academic Press, 1983.

[168] Ottewill, A. C. and Wardell, B. "Quasilocal contribution to the scalar self-force: nongeodesic motion". *Physical Review D* 79.2, 024031 (2009). arXiv:`0810.1961 [gr-qc]`.

[169] Ottewill, A. C. and Wardell, B. "Transport equation approach to calculations of Hadamard Green functions and non-coincident DeWitt coefficients". *Physical Review D* 84.10, 104039 (2011). arXiv:`0906.0005 [gr-qc]`.

[170] Parker, L. E. "Amplitude of perturbations from inflation" (2007). arXiv:`hep-th/0702216`.

[171] Parker, L. E. "Particle creation in expanding universes". *Physical Review Letters* 21.8 (1968), pp. 562–564.

[172] Parker, L. E. "Quantized fields and particle creation in expanding universes. I". *Physical Review* 183.5 (1969), pp. 1057–1068.

[173] Parker, L. E. and Fulling, S. A. "Adiabatic regularization of the energy-momentum tensor of a quantized field in homogeneous spaces". *Physical Review D* 9.2 (1974), pp. 341–354.

[174] Penzias, A. A. and Wilson, R. W. "A measurement of excess antenna temperature at 4080 Mc/s." *Astrophysical Journal* 142 (1965), pp. 419–421.

[175] Pfenning, M. J. "Quantization of the Maxwell field in curved spacetimes of arbitrary dimension". *Classical and Quantum Gravity* 26, 135017 (2009). arXiv:`0902.4887 [math-ph]`.

[176] Pietsch, A. *Nuclear Locally Convex Spaces*. Ergebnisse der Mathematik und ihrer Grenzgebiete 66. Springer, 1972.

[177] Pinamonti, N. "Conformal generally covariant quantum field theory: the scalar field and its Wick products". *Communications in Mathematical Physics* 288.3 (2009), pp. 1117–1135. arXiv:`0806.0803 [math-ph]`.

[178] Pinamonti, N. "On the initial conditions and solutions of the semiclassical Einstein equations in a cosmological scenario". *Communications in Mathematical Physics* 305.3 (2011), pp. 563–604. arXiv:`1001.0864 [gr-qc]`.

[179] Poisson, E., Pound, A., and Vega, I. "The motion of point particles in curved spacetime". *Living Reviews in Relativity* 14.7 (2011).

[180] Powers, R. T. "Self-adjoint algebras of unbounded operators". *Communications in Mathematical Physics* 21.2 (1971), pp. 85–124.

[181] Radzikowski, M. J. "Micro-local approach to the Hadamard condition in quantum field theory on curved space-time". *Communications in Mathematical Physics* 179.3 (1996), pp. 529–553.

[182] Radzikowski, M. J. and Verch, R. "A local-to-global singularity theorem for quantum field theory on curved space-time". *Communications in Mathematical Physics* 180.1 (1996), pp. 1–22.

[183] Rieper, R. G. and Zeleke, M. "Valleyless sequences". In: *Proceedings of the Thirty-first Southeastern International Conference on Combinatorics, Graph Theory and Computing*. Vol. 145. 2000, pp. 33–42. arXiv:`math/0005180 [math.CO]`.

[184] Roura, A. and Verdaguer, E. "Cosmological perturbations from stochastic gravity". *Physical Review D* 78.6, 064010 (2008). arXiv:0709.1940 [gr-qc].

[185] Roura, A. and Verdaguer, E. "Semiclassical cosmological perturbations generated during inflation". *International Journal of Theoretical Physics* 39.7 (2000), pp. 1831–1839. arXiv:gr-qc/0005023.

[186] Ruse, H. S. "Some theorems in the tensor calculus". *Proceedings of the London Mathematical Society* 31.1 (1930), pp. 225–230.

[187] Ruse, H. S. "Taylor's theorem in the tensor calculus". *Proceedings of the London Mathematical Society* 32.1 (1931), pp. 87–92.

[188] Sachs, R. K. and Wolfe, A. M. "Perturbations of a cosmological model and angular variations of the microwave background". *Astrophysical Journal* 147 (1967), p. 73.

[189] Sahlmann, H. and Verch, R. "Microlocal spectrum condition and Hadamard form for vector-valued quantum fields in curved spacetime". *Reviews in Mathematical Physics* 13.10 (2001), pp. 1203–1246. arXiv:math-ph/0008029.

[190] Sanders, K. "Aspects of locally covariant quantum field theory". PhD thesis. University of York, England, 2008. arXiv:0809.4828 [math-ph].

[191] Sanders, K. "Equivalence of the (generalised) Hadamard and microlocal spectrum condition for (generalised) free fields in curved spacetime". *Communications in Mathematical Physics* 295.2 (2010), pp. 485–501. arXiv:0903.1021 [math-ph].

[192] Sanders, K. "On the Reeh–Schlieder property in curved spacetime". *Communications in Mathematical Physics* 288.1 (2009), pp. 271–285. arXiv:0801.4676 [math-ph].

[193] Sanders, K. "The locally covariant Dirac field". *Reviews in Mathematical Physics* 22.4 (2010), pp. 381–430. arXiv:0911.1304 [math-ph].

[194] Sanders, K., Dappiaggi, C., and Hack, T.-P. "Electromagnetism, local covariance, the Aharonov–Bohm effect and Gauss' law". *Communications in Mathematical Physics* 328 (2014), pp. 625–667. arXiv:1211.6420 [math-ph].

[195] Schaefer, H. H. *Topological Vector Spaces*. Graduate Texts in Mathematics 3. Springer, 1971.

[196] Schlemmer, J. "Local thermal equilibrium on cosmological spacetimes". PhD thesis. Universität Leipzig, 2010.

[197] Schlemmer, J. and Verch, R. "Local thermal equilibrium states and quantum energy inequalities". *Annales Henri Poincaré* 9.5 (2008), pp. 945–978. arXiv:0802.2151 [gr-qc].

[198] Schomblond, C. and Spindel, P. "Conditions d'unicité pour le propagateur $\Delta^1(x, y)$ du champ scalaire dans l'univers de deSitter." *Annales Henri Poincaré* 25.1 (1976), pp. 67–78.

[199] Sebestyén, Z. "Every c^*-seminorm is automatically submultiplicative". *Periodica Mathematica Hungarica* 10.1 (1979), pp. 1–8.

[200] Slawny, J. "On factor representations and the c^*-algebra of canonical commutation relations". *Communications in Mathematical Physics* 24.2 (1972), pp. 151–170.

[201] Starobinsky, A. A. "A new type of isotropic cosmological models without singularity". *Physics Letters B* 91.1 (1980), pp. 99–102.

[202] Starobinsky, A. A. "Dynamics of phase transition in the new inflationary universe scenario and generation of perturbations". *Physics Letters B* 117.3–4 (1982), pp. 175–178.

[203] Stewart, J. M. "Perturbations of Friedmann–Robertson–Walker cosmological models". *Classical and Quantum Gravity* 7.7 (1990), pp. 1169–1180.

[204] Stewart, J. M. and Walker, M. "Perturbations of Space-Times in General Relativity". *Royal Society of London Proceedings Series A* 341.1624 (1974), pp. 49–74.

[205] Stoeger, W. R., Maartens, R., and Ellis, G. F. R. "Proving almost-homogeneity of the universe: an almost Ehlers–Geren–Sachs theorem". *Astrophysical Journal* 443 (1995), pp. 1–5.

[206] Streater, R. F. and Wightman, A. S. *PCT, Spin and Statistics, and All That*. The Mathematical Physics Monograph Series. W. A. Benjamin, 1964.

[207] Strohmaier, A. "Microlocal analysis. Concepts and mathematical foundations". In: *Quantum Field Theory on Curved Spacetimes*. Ed. by Bär, C. and Fredenhagen, K. Lecture Notes in Physics 786. Springer, 2009. Chap. 4, pp. 85–127.

[208] Synge, J. L. "A characteristic function in Riemannian space and its application to the solution of geodesic triangles". *Proceedings of the London Mathematical Society* 32.1 (1931), pp. 241–258.

[209] Synge, J. L. *Relativity: The General Theory*. North-Holland, 1960.

[210] Them, K. and Brum, M. "States of low energy in homogeneous and inhomogeneous expanding spacetimes". *Classical and Quantum Gravity* 30.23, 235035 (2013). arXiv:1302.3174 [gr-qc].

[211] Treves, F. *Topological Vector Spaces, Distributions and Kernels*. Pure and Applied Mathematics 25. Academic Press, 1967.

[212] Tsagas, C. G., Challinor, A., and Maartens, R. "Relativistic cosmology and large-scale structure". *Physics Reports* 465.2–3 (2008), pp. 61–147. arXiv:0705.4397 [astro-ph].

[213] Uhlmann, A. "Über die Definition der Quantenfelder nach Wightman und Haag". *Wissenschaftliche Zeitschrift der Karl-Marx-Univversität Leipzig* 11.2 (1962), pp. 213–217.

[214] Unruh, W. G. "Notes on black-hole evaporation". *Physical Review D* 14.4 (1976), pp. 870–892.

[215] Verch, R. "A spin-statistics theorem for quantum fields on curved spacetime manifolds in a generally covariant framework". *Communications in Mathematical Physics* 223.2 (2001), pp. 261–288. arXiv:math-ph/0102035.

[216] Wald, R. M. *General Relativity*. University of Chicago Press, 1984.

[217] Wald, R. M. *Quantum Field Theory in Curved Spacetime and Black Hole Thermodynamics*. University of Chicago Press, 1994.

[218] Wald, R. M. "The back reaction effect in particle creation in curved spacetime". *Communications in Mathematical Physics* 54.1 (1977), pp. 1–19.

[219] Wald, R. M. "Trace anomaly of a conformally invariant quantum field in curved spacetime". *Physical Review D* 17.6 (1978), pp. 1477–1484.

[220] Wrochna, M. "Singularities of two-point functions in quantum field theory". PhD thesis. Universität Göttingen, 2013.

[221] Zschoche, J. "The Chaplygin gas equation of state for the quantized free scalar field on cosmological spacetimes". *Annales Henri Poincaré* 15.7 (2014), pp. 1285–1325. arXiv:1303.4992 [gr-qc].

Index