27

Global dynamics of the universe

Thesis presented for the degree of Doctor of Philosophy

at the Department of Mathematics and Applied Mathematics

of the University of Cape Town by

Jelle Pieter Boersma

Supervisor: George F. R. Ellis

Contents

Prologue

In this thesis we consider four different topics in the field of cosmology, namely, black hole topology, the averaging problem, the effect of surface terms on the dynamics of classical and quantum fields, and the generation of an open universe through inflation with random initial conditions. It should be mentioned that while the research for this thesis was being done, no large effort was made to pursue a single theme. One reason for the diversity of the topics in this thesis is that the results which came out of this research were not always the results which were expected to be found when the investigation was started. Another reason for looking at several topics is simply that once a problem has been solved, then it is natural to move on to another problem which has not yet been solved. For those readers who value that a thesis is centered around a single unifying theme, let me mention that each of the four topics in this thesis are indeed related. Namely, each topic which we discuss focuses on an aspect of the global dynamics of the universe, in a situation where this is non-trivially different from the local dynamics. The non-trivial relation between global and local dynamics is rarely addressed in cosmology. Partially this is because of the difficulties which arise when one considers a realistic universe with infinitely many coupled degrees of freedom. Hence, it is a common practice to rely on simplifications which reduce the number of degrees of freedom, or the couplings between them. Further, there are few direct observations which probe the large-scale dynamics of the universe, or none at all, depending on the length scale and the type of cosmological model which one considers. As a consequence, there is a considerable freedom in choosing *a priori* assumptions or simplifications in the field of cosmology, without being able to falsify the validity thereof. For instance, when we analyse the relation between field perturbations at spatial infinity and perturbations here and now, we assume that quantum field theory, as we know it, is valid everywhere between here and spatial infinity. Although one cannot avoid making certain fundamental assumptions, the type of simplifications which are addopted in a calculation plays a less fundamental role. It is the objective of this thesis to improve our understanding of the large scale dynamics of the universe by showing rigorously what one can and what one cannot derive from certain fundamental assumptions. Interestingly, our results are often quite different from the results which are based on the same assumptions, but which involve certain commonly made simplifications as well. This thesis is structured as follows.

In the first chapter it is shown how different sections of the Kruskal geometry can be identified in a way which preserves time-orientability of the spacetime. The existence of topologically different but locally identical solutions of Einstein's equations is well known, and not surprising considering the differential structure of these equations. We also discuss the occurrence of Hawking radiation in topologically different black-hole geometries. Furthermore, we study the relation between black-hole solutions and circular cosmic strings. Assuming the existence of circular

cosmic strings with deficit angle ranging between 0 and 2π, we are able to construct a class of non-trivial vacuum solutions with properties similar to black-hole solutions but with a more complicated topology.

In the second chapter of this thesis we focus on the averaging problem in cosmology. The averaging problem occurs when one attempts to model a realistic inhomogeneous universe by a more symmetric model. Although averaging is often implied when studying realistic cosmological models, a rigorous treatment of averaging in cosmology appears to be surprisingly difficult. One difficulty which occurs when one tries to specify an averaging procedure is related to the large number of unphysical degrees of freedom which are present in the problem, namely, the coordinate freedom and the gauge freedom. The coordinate freedom manifests itself when one tries to evaluate the average of tensorial quantities, since the components of a tensor depend on the local choice of a frame. One may attempt to avoid this problem by specifying a local frame and evaluating some kind of average for each component separately. However, since there is no choice of frame which is preferred for physical reasons, this gives rise to a considerable amount of ambiguity. When one follows a perturbative approach, there is an additional freedom of choosing a gauge, which makes it ambiguous what one means by a perturbation of a physical quantity, even when this quantity does not depend on the local choice of frame. By specifying a choice of gauge it becomes well defined what one means by a perturbation, but once again no choice of gauge seems to be preferred for physical reasons. In addition to these problems, there is an inherent ambiguity which is related to the freedom in choosing an averaging operation. Since there is generally more than one choice of averaging operation which is mathematically consistent, one needs to impose additional constraints which restrict the freedom of choosing an averaging operation. However, one would like to do so on the basis of a minimal set of assumptions. It is shown that each of these problems can be resolved in the case where perturbations theory can be applied. We use our results to calculate the lowest order non-trivial correction to the expansion of the observable universe, which is due to the fact that averaging does not commute with evaluating the (nonlinear) Einstein equations.

In the third chapter of this thesis we investigate the relation between surface terms which are evaluated at spatial infinity, and the local dynamics of a scalar field. Starting from the path-integral approach to quantum field theory, it is shown that the contribution of surface terms to the variation of the action functional cannot in general be neglected. The classical field equations can be derived by requiring that the variation of the action vanishes for all field perturbations, and it is shown that a surface term generally contributes a non-trivial source term to the classical field equations. This source term appears to vanish in spatially flat geometries, but it diverges in a spatially open geometries with supercurvature perturbations. Rather surprisingly, it appears that the degrees of freedom of the scalar field which generate surface terms must have zero norm in the space of square integrable field

perturbations. Without restricting these zero-norm degrees of freedom, it follows that the local dynamics of the field are sensitive to details of the spacetime at spatial infinity. The main difficulty which we are confronted with consists of quantifying the zero-norm degrees of freedom. We briefly discuss a strategy for resolving this problem.

In the fourth chapter we discuss different types of inflation. As is well known, the standard idea of inflation provides a simple explanation for the homogeneity of the observed universe. However, it appears to be much less straightforward to reconcile a period of inflation with the observed negative spatial curvature in the universe. Bubble inflation combines these two aspects, but it requires a rather restricted type of potential. After introducing the established ideas of standard inflation and bubble inflation, we focus on the dynamics of bubble spacetimes. It is shown that the often used thin-wall approach is not consistent with the assumption that the stress-energy is generated by a scalar field, although this assumption plays a crucial role in the theory of bubble-dynamics. In order to resolve this problem, we derive a simplified set of equations which describe the exact dynamics of a general spherically symmetric bubble spacetime. We then focus on the question of whether the restrictions on the shape of the potential, which are essential in the bubble inflation scenario, are necessary in order to explain the generation of negative spatial curvature during inflation. By studying the most generic situation where constant-scalarfield hypersurfaces make a transition from being spacelike to being timelike, it is shown that negative spatial curvature is generated under conditions which are more generic than the conditions which are generally assumed.

The results which are presented in this thesis have been obtained through independent research, which was conducted by the author on an individual basis. The contents of the first three chapters have been published, [1] - [3], excluding the third section of the first chapter, which was added recently. The contents of the last chapter are currently being prepared for submission. None of the results which are obtained in this thesis have, to the best of my knowledge, been published elsewhere, or the original work has been cited.

References

[1] J. Boersma, Phys. Rev. **D 55**, 2174 (1997).

[2] J. Boersma, Phys. Rev. **D 57**, 1790 (1998).

[3] J. Boersma, Phys. Rev. **D 60**, page number not yet known (1999).

Conventions

Throughout this thesis we adopt the convention that greek indices run from 0 to 3, while latin indices run from 1 to 3. Summation over upper and lower indices is implied, unless it is stated otherwise in the text. We use the metric signature $(-+++)$, and the velocity of light is set equal to one.

1 Black hole topology

Abstract

It is shown how different sections of the Kruskal geometry can be identified in a way which preserves the time-orientability of the spacetime. The geometry which is obtained has the interesting property that it describes an eternal black hole, but the extra asymptotically flat section which occurs in the Kruskal geometry is absent. We briefly discuss the observations of classical observers, and the emission of Hawking radiation in this geometry. Further, we construct an exact solution of Einstein's equations which describes a circular cosmic string, and we show that topologically nontrivial solutions of the homogeneous Einstein equations by identifying sections in different circular cosmic string spacetimes.

1.1 Identifications of the Kruskal Manifold

In this chapter we consider the geometry which describes an eternal spherically symmetric black hole, with vanishing electrical charge and angular momentum. The geometry of the region exterior to the horizon of this type of black hole was determined in the year 1916 by Schwarzschild [1], and the maximal analytic extension of this geometry was found 46 years later by Kruskal [2] and independently by Szekeres [3]. From now on we denote this geometry by $\mathcal{M}$, and the line element on $\mathcal{M}$ is given by

$$\mathrm{d}s^2 = \Omega(u,v)(-\mathrm{d}v^2 + \mathrm{d}u^2) \ + r^2(\mathrm{d}\theta^2 + \sin^2\theta \mathrm{d}\phi^2), \tag{1}$$

where

$$\Omega(u,v) := 32 r^{-1} M^3 e^{-r/2M}, \tag{2}$$

and M denotes the mass of the black hole, and the radial coordinate r is related to the coordinates u and v by,

$$(\frac{r}{2M} - 1)e^{r/2M} = u^2 - v^2. \tag{3}$$

As is clear from equation (2), the metric component Ω satisfies the relation $\Omega(u,v) = \Omega(\pm u, \pm v)$. From now on we will refer to the coordinates $\{u, v, \theta, \phi\}$ as 'Kruskal coordinates'.

A peculiar property of $\mathcal{M}$ is the existence of an extra asymptotically flat universe (see Fig. 1). One may question whether there exists a way to identify the asymptotically flat sections II and IV in $\mathcal{M}$ while preserving the time orientability and regularity of the metric.

Let J be an isometry of $\mathcal{M}$, which has no fixed points, and which satisfies the condition that its square is the identity. By identifying points x and Jx on $\mathcal{M}$, we obtain a quotient manifold $\mathcal{M}/J$ which is nonsingular. Note that the condition that J has no fixed points guarantees that about every two points with coordinates x and Jx in $\mathcal{M}$ there one can find two non-intersecting open environments $O(x)$ and $O(Jx)$. The condition that J is an isometry of $\mathcal{M}$ enforces that the geometry in the two open patches $O(x)$ and $O(Jx)$ is the same. Hence, by identifying points with coordinates x and Jx on $\mathcal{M}$ one obtains a single valued metric, which is *locally* the same as the metric at the points with coordinates x or Jx. The condition that the square of J is the identity is not essential, and one could consider isometries J which have no fixed points and for which $J^n (n \in \mathbf{N}, n > 2)$ equals the identity, or for which only J^0 is equal to the identity. In the case where $J^n (n \in \mathbf{N}, n > 2)$ equals the identity on some type of manifold, then it is easy to show the identification of points with coordinates x and Jx amounts to identifying sets of n different points on this manifold.

Four possible choices for J with the desired properties are

$$J : (v, u, \theta, \phi) \rightarrow \ (s_1 v, s_2 u, \pi - \theta, \phi + \pi), \tag{4}$$

where $s_1, s_2 = \pm 1$. The absence of fixed points for J, as is defined by expression (4), can be easily shown. First observe that a fixed point p with coordinates (v, u, θ, ϕ) must satisfy the equation $p = Jp$, which consists of one condition on each of the four coordinates. If the coordinates u and v satisfy the conditions $v = s_1 v$ and $u = s_2 u$, then the angular coordinates θ and ϕ still have to satisfy the conditions $\theta = \pi - \theta$ and $\phi = \phi + \pi$. For $\theta \in (0, \pi)$ the ϕ-coordinate is non-degenerate, and there are no points for which the ϕ coordinate satisfies the condition $\phi = \phi + \pi$.

For $\theta = 0$ or $\theta = \pi$, the ϕ coordinate degenerates, in the sense that two points with coordinates ϕ and $\phi + \pi$ do not need to be distinct, but in this case the condition $\theta = \pi - \theta$ prevents a solution of the equation $p = Jp$. The identification where $s_1, s_2 = +1$ leads to a black hole with topology $\mathbf{R}^2 \times \mathbf{P}^2$, where $\mathbf{P}^2$ denotes the projected sphere, which is obtained from the two-sphere $\mathbf{S}^2$ by identifying antipodal points. The identification where $s_1 = s_2 = -1$ has been studied extensively in the literature (see [4] - [7]). Since J is time orientation reversing in this case, the quotient manifold $\mathcal{M}/J$ will be non-time-orientable. Namely, a spacetime is time-orientable if and only if this spacetime admits a continuous and everywhere non-vanishing timelike vector field. As an example, the manifold $\mathcal{M}$ is time-orientable, since the vector field $\partial/\partial v$ satisfies the above criterion. We can illustrate the non-time-orientability of $\mathcal{M}/J$ in the case where $s_1 = s_2 = -1$ by the following argument. Let us note that one can define a time-orientation on the section of the quotient manifold $\mathcal{M}/J$ for which $v = 0, u > 0$, by choosing an arrow of time in the direction of the vector field $s\partial/\partial v$, where s denotes an arbitrary sign. One can easily verify that other choices for an arrow of time do not exist, without the vector field vanishing or changing discontinuously somewhere on the section $v = 0, u > 0$.

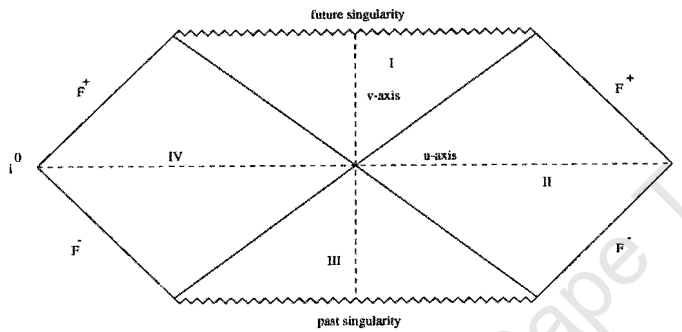

Figure 1: Conformal diagram of the Kruskal manifold $\mathcal{M}$: each point in the diagram represents a two-sphere.

Considering the spacetime $\mathcal{M}$ before identifying points x and Jx, the former choice of an arrow of time points in the direction of $s\partial/\partial v$ on the section of $\mathcal{M}$ for which $v = 0, u > 0$, while it points in the direction of $-s\partial/\partial v$ on the section of $\mathcal{M}$ for which $v = 0, u < 0$ (the minus sign arises from the requirement that a single arrow of time is defined on $\mathcal{M}/J$ after identifying points x and Jx on $\mathcal{M}$). Continuity requires that this vector field must vanish at the two-sphere where $v = 0$ and $u = 0$, and hence a vector field which defines a time-orientation on the section of $\mathcal{M}/J$ for which $u > 0$, cannot be used to define a time-orientation on the entire manifold $\mathcal{M}/J$.

The identification where $s_1 = +1$ and $s_2 = -1$ is time orientation, as well as space orientation, preserving, and it identifies the asymptotically flat regions II and IV. Note, however, that the identification of points x and Jx on $\mathcal{M}$ breaks the symmetry of this spacetime with respect to Schwarzschild time translations, since J does not commute with the generator of Schwarzschild time translations T. The breaking of the Schwarzschild time translation symmetry by identifying points x and Jx on $\mathcal{M}$ manifests itself through a change of the topology of the two-spheres at the line $u = 0$ to $\mathbf{S}^2/J = \mathbf{P}^2$. It appears that the identification of x and Jx with $s_1 = +1$ and $s_2 = -1$ is therefore regarded as unphysical in [4] (i.e., as Israel says, '... any physically meaningful identification must be invariant under the group of transformations which corresponds to time translations $t \to t + const$'), and in

[5] - [7], the possibility of identifying points according to a coordinate-dependent prescription as given here has not been considered.

However, there seems to be no reason why the identification J should preserve the global symmetries of the spacetime in order to be physically meaningful. The source of confusion seems to be that although the prescription of identifying points x and Jx on $\mathcal{M}$ singles out a preferred coordinate system, since the two-spheres for which $u = 0$ and v=constant acquire the topology $\mathbf{P}^2$, it is of no physical relevance in which coordinate system one identifies points x and Jx. Indeed, provided the identification is well defined in one coordinate system, the time translation invariance of $\mathcal{M}$ ensures that the identification of points $\tilde{x}$ and $J\tilde{x}$ in any time-translated coordinate system $\tilde{x} = Tx$ gives rise to the same quotient spacetime $\mathcal{M}/J$. This observation implies that in terms of the non-time-translated coordinate system, all identifications of points x and $T^{-1}JTx$ on $\mathcal{M}$ give rise to the same quotient spacetime $\mathcal{M}/J$. This result shows the uniqueness of the quotient spacetime $\mathcal{M}/J$ (see Fig. 2) even though it was obtained from $\mathcal{M}$ in a way which is not manifestly independent of the choice of coordinates. The physical significance of the breaking of global time-translation invariance by identifying points x and Jx will be discussed in the following. Further, let us note that the relation between global and local symmetries has been considered in a general mathematical context in [8].

Note that the breaking of the T invariance occurs in regions I and III, while region II, which is identified with region IV in $\mathcal{M}$, remains invariant under T.

One might ask oneself the question whether a civilization which discovers an eternal black hole will be able to tell whether this black hole is of type $\mathcal{M}$ or $\mathcal{M}/J$. In order to answer this question, let us consider the observations of classical observers in a spacetime which is either of type $\mathcal{M}$ or $\mathcal{M}/J$. Note that for $u > 0$ the spacetimes $\mathcal{M}$ and $\mathcal{M}/J$ are identical, and therefore an observer who remains entirely within the region II has no means to tell whether or not this region of spacetime belongs to $\mathcal{M}$ or $\mathcal{M}/J$. An observer in region II could send an explorer into region I, and stay in region II himself, but this will yield him little wisdom since no information about the findings of the explorer in region I can be transmitted to region II. Finally the observer could decide to cross the horizon himself. If he had arranged for some other object or light signal to go into the black hole, starting from the point with antipodal angular coordinates with respect to the observer, then the observer would have a chance to encounter this object or light signal after one of them has passed the line $u = 0$ in Fig. 2 (such a trajectory is marked A in Fig. 2; note that observers which cross the line $u = 0$ re-emerge with antipodal angular coordinates). If this would happen, then the observer would know that his spacetime is of the type $\mathcal{M}/J$. However, an observer who crosses the horizon could be so unfortunate as to start too late in order to reach the object or light ray which has gone into the black hole with antipodal angular coordinates, and in this case the observer will end up in the singularity without knowing whether he lives in $\mathcal{M}$ or $\mathcal{M}/J$ (such a trajectory is marked B in Fig. 2). This would happen in any case

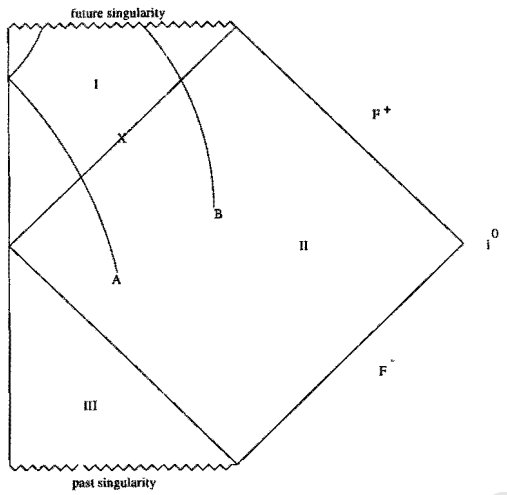

Figure 2: Conformal diagram of the quotient manifold $\mathcal{M}/J$: each point in the diagram represents a two-sphere $\mathbf{S}^2$, except for points at the left boundary of the diagram, which represent a sphere $\mathbf{P}^2$.

if the observer or the object or light signal do not cross the horizon before the event which is marked with an X in Fig. 2, since in this case both the observer and the object or light signal will only have the region $u > 0$ in their future light cone, for which the spacetimes $\mathcal{M}$ and $\mathcal{M}/J$ are identical. The fact that there is a latest time before which the observer and the light signal must have crossed the horizon in $\mathcal{M}/J$ in order to find out whether they live in $\mathcal{M}$ or $\mathcal{M}/J$ illustrates that the time translation symmetry which is present in $\mathcal{M}$ is broken at a global level in $\mathcal{M}/J$. Notice that there is no way by which the observer which stays in region II of $\mathcal{M}/J$ can find out whether he is still in time to cross the horizon before the point X, without entering region I. This is due to the fact that region II remains invariant under the time translations generated by T. We therefore established that classical observers who live in the asymptotically flat region outside the black hole horizon, have no secure means to probe the topology of the black hole. In the following section, we will show that risky experiments where the observer traverses the horizon are not necessary in order to find out whether an eternal black hole is of type $\mathcal{M}$ or $\mathcal{M}/J$. This is due to the quantum effects in the spacetimes $\mathcal{M}$ and $\mathcal{M}/J$.

1.2 Quantum effects and black hole topology

Before we discuss the quantum effects in the spacetime $\mathcal{M}/J$, I will briefly summarize some ideas behind Hawking's derivation of particle creation in the spacetime $\mathcal{M}$. As was shown by Hawking in [9], an observer in region II of the Kruskal manifold measures quanta of a quantum field ϕ when the quantum field is in the Hartle-Hawking vacuum state, for which the fields are analytic in terms of the Kruskal coordinates. The number of particles which is measured by an observer in region II in $\mathcal{M}$ can be derived by expanding the field operator corresponding to the quantum field ϕ in terms of a complete set of solutions of the field equation, which have the form

$$\phi^{\pm}_{\omega lm}(u,v,\phi,\theta) = r^{-1} f^{\pm}_{\omega,l,m}(u,v) Y_{lm}(\phi,\theta), \tag{5}$$

where we have factorized the solutions in terms of a u,v-dependent part $f_{\omega,l,m}(u,v)$ and the spherical harmonics $Y_{lm}(\phi,\theta)$. Indeed, since the spherical harmonics $Y_{lm}(\phi,\theta)$ are known to be complete in the space of square integrable functions on the two-sphere, it follows that any square integrable function ϕ on $\mathcal{M}$ can be written as the sum over l and m of a spherical harmonic Y_{lm} times a coefficient which depends on u,v,l and m. We may choose these coefficients to be orthonormal with respect to integration over $\mathcal{M}$, while the existence of a complete set of modes on $\mathcal{M}$ follows from the Hilbert-space property of the space of square integrable functions (see, *e.g.*, [10]). Without loss of generality, we may introduce a splitting of the u,v-dependent part of the ϕ mode in terms of symmetric and antisymmetric functions, which satisfy the condition $f^{\pm}_{\omega,l,m}(u,v) = \pm f^{\pm}_{\omega,l,m}(-u,v)$.

There exists a natural decomposition of the ϕ modes, given by expression (5), in terms of solutions which have positive or negative frequency with respect to the Schwarzschild time parameter. We may also consider solutions of the form (5) which oscillate as a function of the Kruskal time parameter v, and unlike the solutions which oscillate as a function of the Schwarzschild time parameter, these solution are analytic at the horizon where $u = \pm v$.

The field quantization associated with the decomposition of ϕ modes in terms of $\pm$ frequency solutions with respect to Schwarzschild time parameter, and the field quantization which is based on solutions which are an analytic function of the Kruskal time parameter v, appear to be inequivalent. With these different field quantizations there are associated inequivalent sets of creation and annihilation operators and vacuum states. The transformation which relates these sets of creation and annihilation operators can be determined systematically, and is called a Bogoliubov transformation [10]. This allows one to derive an expression for the expectation value of ϕ particles, in the state represented by the field mode (5), as measured by a static observer outside the black hole:

$$\langle 0|N_{\omega lm}|0\rangle = |t|^2 \frac{1}{e^{2\pi\omega/\kappa} - 1}. \tag{6}$$

In this expression $|t|^2$ is the absorption cross section of the black hole for the mode $\phi_{\omega lm}$, κ is the surface gravity of the black hole, and the vacuum state $\langle 0|$ is taken to be the vacuum state which is natural in terms of the Kruskal-time quantization. Now we may ask ourselves the question of what would change in the expression of the expectation value of ϕ particles if we had performed our calculation in the spacetime $\mathcal{M}/J$ instead of the spacetime $\mathcal{M}$. Fortunately, we do not need to start from the beginning in order to determine the particle creation rate in $\mathcal{M}/J$ if we make use of the observation that the spacetime $\mathcal{M}$ is a covering spacetime of $\mathcal{M}/J$. Instead of considering the field theory on $\mathcal{M}/J$, one can therefore equivalently consider the field theory on the covering spacetime $\mathcal{M}$, where the fields are subject to the condition

$$\phi(x) = \phi(Jx). \tag{7}$$

The latter condition ensures that we consider only those solutions on $\mathcal{M}$ which give rise to a single-valued solution on $\mathcal{M}/J$. For the symmetric ϕ modes $\phi^{+}_{\omega lm}$ we find that condition (7) implies the condition

$$Y_{lm}(\phi, \theta) = Y_{lm}(\pi + \phi, \pi - \theta) \tag{8}$$

on the spherical harmonics Y_{lm}. Condition (8) is satisfied if and only if the quantum number m, which takes values in $[-l, -l+1, ..., l-1, l]$, is restricted to even numbers. Similarly, for the antisymmetric ϕ modes $\phi^{-}_{\omega lm}$ we find that condition (7) reduces to the condition

$$Y_{lm}(\phi, \theta) = -Y_{lm}(\pi + \phi, \pi - \theta) \tag{9}$$

on the spherical harmonics Y_{lm}. Condition (9) is satisfied if and only if the quantum number m is restricted to odd numbers. Apart from the restriction on the ϕ modes which arise from conditions (7) and (9), the derivation of the Hawking effect on the covering spacetime $\mathcal{M}/J$ is identical to the derivation of the Hawking effect on the Kruskal manifold $\mathcal{M}$. The expectation value for field quanta in the quantum state with energy ω and quantum numbers l, m on $\mathcal{M}/J$ will therefore be once more given by expression (6), but this time the extra condition {m must be even } applies for the symmetric field modes $\phi^{+}_{\omega lm}$, while the condition {m must be odd } applies for the antisymmetric modes $\phi^{-}_{\omega lm}$. However, notice that a ϕ quantum on $\mathcal{M}/J$ has twice as much chance of being detected as an observer in region II, as would be the case if this observer lived in $\mathcal{M}$ instead of $\mathcal{M}/J$. This is due to the fact that the wave functions $\phi^{\pm}_{\omega lm}$ in $\mathcal{M}/J$ are normalized by integrating the probability density for this wave function over a Cauchy surface in $\mathcal{M}/J$ (*e.g.*, $v = 0, u \geq 0$). It is easy to show that the probability density for normalized quanta in $\mathcal{M}/J$ must exactly double the probability density for normalized quanta in $\mathcal{M}$.

1.3 Black hole cosmic strings

In this section we discuss an identification of points in the spacetime $\mathcal{M}$ which changes both the topology and the matter content of the spacetime. Let us

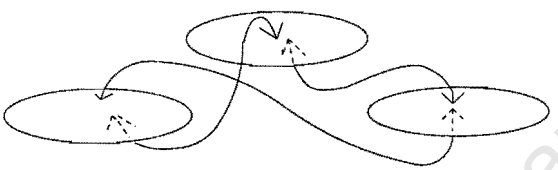

Figure 3: A sourceless solution of Einstein's equation is obtained by identifying the lower and upper surface of three discs which are bounded by three circular cosmic strings with deficit angle $\frac{4}{3}\pi$.

therefore define the isometry J by its action on points p in $\mathcal{M}$ with coordinates $\{u, v, \theta, \phi\}$,

$$J : (v, u, \theta, \phi) \rightarrow \ (v, -u, \pi - \theta, \phi). \tag{10}$$

By working out the condition $Jx = J$ for points x with coordinates $\{v, u, \theta, \phi\}$, it follows directly that the isometry J has a fixed point if and only if $u = 0$ and $\theta = \pi/2$. The family of fixed points of J, which we denote by the symbol p, is therefore described by two parameters v and ϕ, where the time-coordinate v takes values in a compact interval in $\mathbf{R}$ (this follows from expression (3), where we use that the radial coordinate r takes values in a compact interval), while the angular coordinate ϕ is cyclic with a period equal to 2π. One can therefore visualize the collection of fixed points p as a $1 + 1$ dimensional worldsheet of a circular string which evolves in time along the line $u = 0$. Let us now consider the identification of points x and Jx on the subspace $\mathcal{M}^*$ which consists of those points in $\mathcal{M}$ which are not fixed points of J. Indeed, since J is an isometry of $\mathcal{M}^*$ which has by definition no fixed points, it follows that the quotient manifold $\mathcal{M}^*/J$ is well defined, and has locally the same properties as $\mathcal{M}^*$. It is clear from expression (10) that J maps points in one asymptotically flat region in $\mathcal{M}$ onto the other asymptotically flat region in $\mathcal{M}$, such that the quotient manifold $\mathcal{M}^*/J$ has only one asymptotically flat region.

Note also that the identification of points x and Jx in $\mathcal{M}$ does change the intrinsic geometry of $\mathcal{M}$ at the fixed points p of the isometry J. The geometrical structure of the spacetime at the points p is easily understood by considering a closed circle about the string of points p in $\mathcal{M}$. Let us therefore consider a specific point p_1 on the two-dimensional subset of $\mathcal{M}$ which consists of points p, and we define $\Sigma(p_1)$ to be a two-dimensional plane through p_1 which is tangent to the vectors $\partial/\partial v$ and $\partial/\partial\phi$. Since $\mathcal{M}$ is locally flat everywhere outside the two singularities, it follows that the two-plane $\Sigma(p_1)$ can be chosen so that it is a locally flat two-dimensional plane through p_1, which is coordinatized by the coordinates $\{\theta-\pi/2, u\}$. The operation of J on $\Sigma(p_1)$ maps points with coordinates $\{\theta-\pi/2, u\}$ onto points with coordinates $\{-\theta+\pi/2, -u\}$, so that the identification of points x and Jx at $\Sigma(p_1)$ generates a conical singularity at p_1, which is characterized by a deficit angle π. Hence, by identifying points in $\mathcal{M}$, which is a homogeneous solution of Einstein's equation, we have obtained a solution of the inhomogeneous Einstein equations describing a circular cosmic string. It is of interest to note that the same method can be applied vice-versa, *i.e.*, by 'cutting and pasting' sections in disconnected spacetimes which describe identical circular cosmic strings, one can obtain a solution of the homogeneous (*i.e.*, sourceless) Einstein equations. The case where one obtains the solution $\mathcal{M}$ from two identical circular cosmic string solutions is fairly trivial, but not of much interest, since the solution $\mathcal{M}$ is already known. A more interesting situation occurs if one considers a solution of Einstein's equations which describes a circular cosmic string with a deficit angle equal to $\frac{4}{3}\pi$. We have not proven the existence of this type of solution, but it seems rather plausible that this type of cosmic string solutions do indeed exist. Indeed, the geometry which describes a straight cosmic string with a deficit angle in the range $(0, 2\pi)$ can be constructed in a simply fashion by identifying points at the two half-planes which form the boundary of a wedge in Minkowski spacetime. For this type of string geometry the deficit angle Δ is related to the energy per length unit μ by the simple expression (see, *e.g.*, [11]),

$$\Delta = 8\pi G\mu. \tag{11}$$

Although a circular string geometry is globally rather different from a straight string geometry, one expects that the difference between these two types of geometry becomes important only at length scales which are of the order of the circumference of the circular string. Making this assumption, one may then consider the limit where the circumference of the circular string becomes arbitrarily large, and it follows that the circular and the straight string geometry converge to the same limit when considered at a fixed length scale. Indeed, let us note that the circumference of the circular string is the only natural length scale which is present in the problem, and hence scaling the circumference of the string is equivalent to redefining our measure of length. These observations suggest that a circular cosmic string has the same local properties as the straight cosmic string, and in particular

one expects that the relation between the deficit angle and the mass per unit length is the same in both geometries.

Let us now consider three identical circular string geometries, with deficit angle $\frac{4}{3}\pi$. Note that each of the circular strings forms the boundary of an open two-dimensional flat disc. We may then identify points at the top of the first disc with points at the bottom of the second disc, and we continue to do so cyclically, as is shown in figure 1.2. It is clear that this identification does not change the local structure of the spacetime, as long as the geometry at each of the three discs is identical (more formally, the intrinsic and the extrinsic curvature of the two sides of the disc which are identified must be identical, in order to assure that the spacetime which is obtained by this identification is still a solution of the homogeneous Einstein equations; see also the discussion in section 4.4 in chapter 4). Although the local structure of the spacetime does not change by identifying points according to this prescription, the *global* structure of the spacetime, as well as the geometry at the cosmic string do change. In order to quantify the effect of this cutting and pasting exercise, it is illuminating to consider a closed loop with an arbitrarily small but constant radius, which circles the cosmic string in the plane $\Sigma(p)$. One may then calculate the fraction of the circumference and the radius of the loop which circles the cosmic string. The fact that the deficit angle is equal to $\frac{4}{3}\pi$ in each of the cosmic string solutions means that this fraction equals to $\frac{2}{3}\pi$. After the identification which we proposed, it is clear that the loop which circles the cosmic string must do so in three different spacetimes subsequently, before it re-arrives at the point from where it started. Hence the deficit angle vanishes for this type of geometry, and it follows that the mass of the string must vanish. In the case where the circumference of the string is time dependent, the string tension must also vanish due to stress-energy conservation, which implies that we are dealing with a sourceless solution of Einstein's equations. The interesting property of this solution is that it seems to have properties similar to a black hole solution, but its topological structure is more complicated in the sense that it connects three different manifolds.

1.4 Conclusions

We conclude with the remark that the two asymptotically flat sections in the Kruskal manifold can be identified while preserving the time orientability of the metric. We found that for classical observers who remain in the asymptotically flat region, the Kruskal manifold $\mathcal{M}$ is indistinguishable from the quotient spacetime $\mathcal{M}/J$, while observers who cross the horizon have an uncertain chance of being able to distinguish between $\mathcal{M}$ and $\mathcal{M}/J$. However, the Hawking effect manifests itself differently on $\mathcal{M}$ and $\mathcal{M}/J$, since additional restrictions apply to the quantum numbers of the ϕ particles which can be emitted by a black hole of type $\mathcal{M}/J$. Namely, for a given frequency ω and an even or odd value of the angular wavenumber m, the black hole of type $\mathcal{M}/J$ radiates only quanta for which the

radial wavefunction is symmetric or antisymmetric respectively. This difference in the spectrum of the Hawking radiation emitted by black holes of type $\mathcal{M}$ and $\mathcal{M}/J$ could possibly be of observational interest if an eternal black hole, or more realistically a primordial black hole, would be discovered. Further, we discussed a relation between black hole solutions and circular cosmic strings. It was shown that topologically nontrivial sourceless solutions of Einstein's equations can be constructed from circular cosmic string solutions.

Additional note: Various aspects of the black hole topology which is constructed in the first section of this chapter have been discussed in two independent papers. One of these papers, by Chamblin and Gibbons [12], was published simultaneously with the contents of the first two sections of this chapter, [13], and it focuses on the pin and the spin structure of the manifold $\mathcal{M}/J$. However, recently an earlier paper by Friedman *et al.* came to my attention, [14], which was apparently also unknown to the authors of [12], and this paper discusses the geometry $\mathcal{M}/J$ in relation to the topological censorship hypothesis.

References

[1] K. Schwarzschild, Sitzber. Deut. Akad. Wiss. Berlin, K1. Math.-Phys. Tech. **21**, 189 (1916).

[2] M. D. Kruskal, Phys. Rev. **119**, 1743 (1960).

[3] G. Szekeres, Publ. Mat. Debrecen **7**, 285 (1960).

[4] W. Israel, Phys. Rev. **143**, 1016 (1966).

[5] G. W. Gibbons, Nucl. Phys. **B 271**, 497 (1985).

[6] F. J. Belinfante, Phys. Lett. **20 A**, 25 (1966).

[7] J. L. Anderson and R. Gautreau, Phys. Lett. **20 A**, 24 (1966).

[8] G. S. Hall, Class. Quant. Grav. Grav. **6**, 157 (1989).

[9] S. W. Hawking, Commun. Math. Phys. **43**, 199 (1975).

[10] R. M. Wald, *Quantum field theory in curved spacetime and black hole thermodynamics* (University of Chicago press, Chicago, 1994).

[11] A. Vilenkin and E. P. S. Shellard, *Cosmic strings and other topological defects* (Cambridge University Press, Cambridge, 1993).

[12] A. Chamblin and G. W. Gibbons, Phys. Rev. **D 55**, 2177 (1997).

[13] J. Boersma, Phys. Rev. **D 55**, pg. 2174 (1997).

[14] J. Friedman *et al.*, Phys. Rev. Lett. **71** 1486 (1993).

2 Averaging in cosmology

Abstract

In this chapter we discuss the effect of local inhomogeneities on the global expansion of nearly Friedmann-Lemaitre-Robertson-Walker (FLRW) universes, in a perturbative setting. We derive the unique linearized averaging operation for metric perturbations from basic assumptions. This averaging operation is used to determine a gauge invariant expression for the backreaction of density inhomogeneities on the global expansion of perturbed FLRW spacetimes. We express our result in terms of observable quantities, and we calculate the effect quantitatively. Since we do not adopt a comoving gauge, our result incorporates the backreaction on the metric which is due to scalar velocity and vorticity perturbations. The results are compared with the results by other authors in this field.

2.1 Introduction

An essential difficulty which occurs when dealing with realistic cosmological models is related to the fact that although the universe seems to be very close to a Friedmann-Lemaître-Robertson-Walker (FLRW) spacetime at length scales of the order of the Hubble radius, the metric and matter content of the universe appears to be highly inhomogeneous at smaller scales. Since the realistic universe, with all its details at length scales small compared to the Hubble radius, is too complicated to handle in most calculations, it seems desirable to extract those physical quantities which describe the large scale structure of the universe. However, when one tries to define an averaging operation for metrics, a number of difficulties occur. One of these difficulties is related to the fact that the Einstein equations are inherently nonlinear, which makes it a nontrivial question to see how the Einstein equations constrain the dynamics of an averaged metric. Another fundamental problem which occurs when one tries to average metrics, is related to the fact that there is generally no direct physical significance in an averaged metric. Although this problem is usually ignored in the literature on averaging, it needs to be addressed before one can extend the discussion on averaging beyond an intuitive level of understanding. The usual approach to averaging (see *e.g.*, [1] -[5]) seems to be that one defines an averaging method, which is chosen on the basis of mathematical elegance or an intuitive notion of smoothness, and then one defines averaged physical quantities by means of the averaging operation which one has chosen. The objection against this approach is that if one calculates, *e.g.*, the averaged expansion of a perturbed FLRW universe, one can obtain virtually any result, by choosing an averaging operation which yields this specific result. In section 2.2

we address this problem, and we derive a generic linearized averaging operation for *metrics*, which satisfies the condition that an unperturbed FLRW spacetime is a *stable fixed point* of the averaging operation. It is shown that this generic linearized averaging operation for metrics can be expressed in terms of the spatial average of the perturbation of the spatial volume and g_{00} in coordinates which are synchronous in the background. In section 2.3 we discuss the gauge problem and the choice of the background spacetime. The averaged constraint equations are explicitly evaluated in section 2.4, and in subsection 2.5 we derive an expression for the correction to averaged expansion due to density perturbations, in terms of the power spectrum of the matter. In subsections 2.6 and 2.7 we discuss the backreaction on the metric due to matter velocity perturbations, and we show that vorticity perturbations may be important in the long wavelength limit. In section 2.8 we calculate the different corrections to the averaged expansion quantitatively by means of the observational data, and we compare our results with the results derived in previous works.

2.2 The averaged metric

The idea of averaging a metric is that one defines a new metric in terms of the former metric, such that the geometry which is described by the new metric is smoother than the geometry which is described by the former metric. The simplest way to average a metric, or any other quantity which transforms as a tensor, is by contracting this tensor with a weighting bi-tensor, and to integrate this expression over the spacetime. In order to illustrate this, let us consider the most general situation where $T_{\mu\nu\ldots}(x)$ is a tensor field over a spacetime $\mathcal{M}$. Further, let $A^{\mu\nu\cdots}_{\alpha\beta\ldots}(x,x')$ be a weighting bi-tensor, where the indices $\alpha\beta\cdots$ transform as a tensor with respect to coordinate transformations at the point x, while the indices $\mu\nu\cdots$ transform as a tensor with respect to coordinate transformations at the point x'. A possible way to average the tensor field T over $\mathcal{M}$, in a way which is consistent with the requirement of coordinate invariance, is obtained by evaluating the integral

$$\langle T_{\alpha\beta\ldots}\rangle = \int_{\mathcal{M}} \mathrm{d}^4x' \; A^{\mu\nu\cdots}_{\alpha\beta\ldots}(x,x') \; T_{\mu\nu\ldots}. \tag{12}$$

Clearly, expression (12) does not describe the most generic averaging procedure for tensors, since we have restricted ourselves to the case where the averaging operation is linear in the argument $T_{\mu\nu\ldots}$. Further, there is a certain amount of freedom involved in the choosing the weighting bi-tensor $A^{\mu\nu\cdots}_{\alpha\beta\ldots}(x,x')$. In order to motivate a certain choice of averaging operation, one will have to invoke physical or philosophical arguments. The freedom to choose different philosophical criteria, which may result in different choices of averaging operation, appears to be a fundamental but rarely addressed ambiguity which underlies the concept of averaging.

In the following we will concentrate on the problem of averaging perturbations of a metric. Starting from the most generic averaging operation for metric pertur-

bations, we derive the generic linearized averaging operation for metrics perturbations, by imposing the condition that an unperturbed FLRW spacetime is a stable fixed point of this averaging operation. We then show that this averaging operation has a universal limit when applied iteratively to perturbed FLRW spacetimes. We determine this limit explicitly, and by using the symmetry of the background, we show that the averaging of the ten components of the metric perturbation $\delta g_{\rho\sigma}$, can be expressed in terms of the spatial average of g_{00} and $\sqrt{g^{(3)}}$ in coordinates which are synchronous in the background. From now on the background FLRW spacetime will be called $\bar{\mathbf{S}}$, while the inhomogeneous spacetime is called $\mathbf{S}$. Furthermore, we assume that $\bar{\mathbf{S}}$ is coordinatized such that t represents the time coordinate which labels the hypersurfaces of homogeneity $\bar{\Sigma}$ in $\bar{\mathbf{S}}$, and $\bar{\Sigma}$ is coordinatized by x^i, where $i \in \{1,2,3\}$. We call a metric $g_{\mu\nu}$ or a metric perturbation $\delta g_{\mu\nu}$ spatially homogeneous and isotropic when there exists at least one coordinate system in which the components of $g_{\mu\nu}$ or $\delta g_{\mu\nu}$ are spatially homogeneous and invariant under spatial rotations.

Let us consider the most general averaging operation $\hat{A}$, which is a functional $\mathcal{F}$ of metric perturbations $\delta g_{\mu\nu}$ about some background solution $\bar{\mathbf{S}}$,

$$\hat{A}\delta g_{\mu\nu}(x) \ = \mathcal{F}_{\mu\nu}[\delta g_{\rho\sigma}(x)]. \tag{13}$$

We *require* the condition that an unperturbed FLRW spacetime, in a gauge where the metric perturbation $\delta g_{\mu\nu}$ is spatially homogeneous and isotropic, is a stable fixed point of the averaging operation $\hat{A}$. This condition states that the averaging operation increases the spatial symmetry of the spacetime on which it works, assuming that this spacetime is sufficiently 'close' to FLRW spacetime, and it defines what we mean by averaging.

It follows directly from this assumption that

$$\mathcal{F}_{\mu\nu}(0) = 0, \tag{14}$$

for all x, since a nonzero value at the right–hand side of equation (14) implies that unperturbed FLRW spacetime with the the same geometry as $\bar{\mathbf{S}}$, in a gauge where $\delta g_{\mu\nu} = 0$ for all x, is not a fixed point of $\hat{A}$, which contradicts our assumption.

The *linear approximation* to the averaging operation (13), is given by

$$\hat{A}^{(1)}\delta g_{\mu\nu}(x) = \int_{\bar{S}} d^4x' \ f^{\rho\sigma}_{\mu\nu} \ (x,x')\delta g_{\rho\sigma}(x'), \tag{15}$$

where the bi-tensor density $f^{\rho\sigma}_{\mu\nu}(x,x')$ is defined as the functional derivative of $\mathcal{F}_{\mu\nu}$ with respect to $\delta g_{\rho\sigma}$, evaluated at the point with coordinates x' in the background, *i.e.*,

$$f^{\rho\sigma}_{\mu\nu}(x,x') := \left.\frac{\delta \mathcal{F}_{\mu\nu}(g,x)}{\delta g_{\rho\sigma}(p)}\right|_{\delta g_{\rho\sigma}=0, p=x'}, \tag{16}$$

and we used condition (14) which states that the zeroth order contribution in the expansion of $\hat{A}$ vanishes.

The condition that unperturbed FLRW spacetime is a *stable* fixed point of the averaging operation $\hat{A}$ implies that the limit

$$\hat{A}^{\infty}\delta g_{\mu\nu} := \lim_{n\to\infty} \hat{A}^{(1)n}\delta g_{\mu\nu} \tag{17}$$

exists, and the quantity $\hat{A}^{\infty}\delta g_{\mu\nu}$ must be spatially homogeneous and isotropic (we used the notation $\hat{A}^{(1)n}$ to denote the n-times repeated operation of $\hat{A}^{(1)}$).

Note that the averaging operation $\hat{A}$ has two aspects; first it changes the geometry of the spacetime on which it works, and second it specifies a correspondence between points in the spacetime $\mathbf{S}$, the averaged spacetime $\hat{A}\mathbf{S}$, and the background $\bar{\mathbf{S}}$. When one only requires that unperturbed FLRW is a stable fixed point of $\hat{A}$, one constrains the way in which $\hat{A}$ changes the geometry of the spacetime on which it works, but one does not constrain the correspondence between points in the spacetimes $\mathbf{S}$, $\hat{A}\mathbf{S}$, and $\bar{\mathbf{S}}$. We constrain this freedom by imposing the *stronger* requirement that unperturbed FLRW spacetime, in a gauge where the metric perturbation $\delta g_{\mu\nu}$ is spatially homogeneous and isotropic, is a stable fixed point of $\hat{A}$. This condition ensures that $\hat{A}$ does not generate 'pure gauge' perturbations when operating on unperturbed FLRW spacetime.

Starting from equation (17), and using the symmetries of the background spacetime $\bar{\mathbf{S}}$, it is shown in appendix 2.B that the averaging operation $\hat{A}^{\infty}$ can be defined in terms of a spatial averaging operation which is universal, *i.e.*,

$$\hat{A}^{\infty}\delta g_{\mu\nu}(t, x^i) = \langle \delta g_{\mu\nu} \rangle (t), \tag{18}$$

where

$$\langle \delta g_{\mu\nu} \rangle (t) = \int_{\bar{\Sigma}(t)} d^3x' \alpha \sqrt{g^{(3)}} \tag{19}$$

$$\times \left[\bar{n}^{\rho}(x')\bar{n}^{\sigma}(x')\, \bar{n}_{\mu}(x)\bar{n}_{\nu}(x) \; + \frac{1}{3}\bar{h}^{\rho\sigma}(x')\bar{h}_{\mu\nu}(x) \right] \; \delta g_{\rho\sigma}(x'),$$

where $\bar{n}^{\rho}$ denotes the future directed unit vector normal to $\bar{\Sigma}$, and $\bar{h}^{\rho\sigma} := g^{B\rho\sigma} + \bar{n}^{\rho}\bar{n}^{\sigma}$ is the projection operator on $\bar{\Sigma}$, and α denotes the distribution which is constant on Σ, and for which the integral over Σ equals 1. Note that $\bar{n}^{\rho}\bar{n}^{\sigma}\delta g_{\rho\sigma}$ equals the perturbation of g_{00} in coordinates which are synchronous in the background (*i.e.*, coordinates for which $g^{B}_{\mu 0} = -\delta^{0}_{\mu}$), while $\bar{h}^{\rho\sigma}\delta g_{\rho\sigma}$ equals the perturbation of the spatial volume element on Σ, to first order. It follows from this observation that the linearized averaging operation for metrics (19), is effectively a spatial averaging operation for scalars, applied to δg_{00} and δg^{i}_{i} in coordinates which are synchronous in the background.

An explicit realization of the spatial averaging operation for a scalar $q(x)$, in the case where $\bar{\Sigma}$ is open, is given by

$$\langle q(x) \rangle = \lim_{\ell\to\infty} \langle q(x) \rangle (\ell) \tag{20}$$

$$:= \lim_{\ell\to\infty} \; N^{-1}(x,\ell) \; \int_{\bar{\Sigma}} d^3x' \sqrt{g^{(3)}} \; q(x')\theta[\ell - \Delta s(x,x')],$$

where $N(x,\ell) := \int_\Sigma d^3x' \sqrt{g^{(3)}}\, \theta[\ell - \Delta s(x,x')]$, and $\Delta s(x,x')$ is a distance measure between points x and x', ℓ is a parameter with the dimension of length, and $\theta(x) = 1(0)$ for $x \geq 0 (x < 0)$. In the case where Σ is closed, $\langle q \rangle$ is defined analogously to expression (20), with $N(x,\ell)$ =volume (Σ).

It is shown in appendix 2.B that the spatial average of a scalar function is invariant under *spatial* gauge transformations, to arbitrary order in the expansion parameter of the gauge transformation.

Notice that the spatial average (20) is only well defined when we make the assumption that perturbations $q(x)$ are sufficiently small, such that the limit $\ell \to \infty$ in equation (20) exists. It should be stressed that this assumption is nontrivial, and it is not automatically satisfied in general cosmological situations, where perturbations are not necessarily bounded in amplitude and length scale. Indeed, since the observable part of our universe is restricted to our past light cone, there is no observational basis for the assumption that our universe is 'close' to FLRW at arbitrary large length scales. The usual way to deal with this situation is that one adopts *a priori* philosophical assumptions, such as the Copernican principle, to choose between different models which satisfy the observational data (see, *e.g.*, [8]). Throughout our calculation, we will adopt a version of the Copernican principle by assuming that perturbations are small enough such that the limit $\ell \to \infty$ in equation (20) exists.

2.3 The gauge problem

As pointed out by Futamase in [1], the observed matter density contrasts are of the order of unity at dimensionless length scales κ of the order of 10^{-2}, where κ denotes the fraction of typical size of the density fluctuation and the Hubble radius $r_H := c/H_0$. A rough estimation of the order of magnitude of the associated Newtonian gravitational potential, which we call ϵ from now on, can be obtained by using the Poisson equation. For density contrasts of the order of unity we find $\epsilon \sim \kappa^2$, which implies a Newtonian potential ϕ of the order of 10^{-4}, suggesting that a perturbative approach might be adequate. At length scales of the order of the Hubble radius, the observable part of the universe appears to be highly homogeneous and isotropic, which motivates our choice for the FLRW metric as a background metric.

Let us first briefly discuss some details concerning the spherical harmonic decomposition of perturbations about a background FLRW spacetime. The FLRW background metric can be written in the form,

$$ds^2 = g^B_{\mu\nu} dx^\mu dx^\nu = a^2(\bar{t})(-d\bar{t}^2 + \eta_{ij} dx^i dx^j), \tag{21}$$

where η_{ij} is the metric tensor for a homogeneous and isotropic three-space with curvature $\mathbf{k}$, and $\bar{t}$ is a *conformally scaled* time parameter. We define the metric

perturbation $h_{\mu\nu}$ by

$$g_{\mu\nu} = g^B_{\mu\nu} + h_{\mu\nu} \quad , \qquad g^{\mu\nu} = g^{B\mu\nu} - h^{\mu\nu}, \tag{22}$$

and since $g^{\mu\rho}g_{\mu\nu} = \delta^\rho_\nu$, we have $h^{\mu\nu} = g^{B\nu\rho}g^{B\mu\sigma}h_{\rho\sigma}$, and $h^\mu_\nu = g^{B\mu\rho}h_{\rho\nu}$, to first order.

Copying Bardeen's notation in [6], we define scalar, vector and tensor spherical harmonics $Q^{(0)}_n$, $Q^{(1)}_{n\,i}$, and $Q^{(2)}_{n\,ij}$, respectively, which satisfy the Helmholtz equations $Q^{(p)|q}_{n\quad|q} + k_n^2 Q^{(p)}_n = 0$, where $p \in \{0,1,2\}$ and $|$ denotes the covariant derivative with respect to g^B_{ij}. The vector harmonics $Q^{(1)}_i$ are divergenceless, while the tensor harmonics $Q^{(2)}_{ij}$ are divergenceless, symmetric, and traceless. We define traceless symmetric scalar harmonics $Q^{(0)}_{n\,ij}$ by

$$Q^{(0)}_{n\,ij} := k_n^{-2} Q^{(0)}_{n|ij} + \frac{1}{3} g^B_{ij} Q^{(0)}_n, \tag{23}$$

and these modes satisfy the equation

$$Q^{(0)ij}_{n\quad|ij} - (k_n^2 - 3\mathbf{k}) Q^{(0)}_n = 0. \tag{24}$$

Further, we define the traceless symmetric vector harmonics $Q^{(1)}_{n\,ij}$ by

$$Q^{(1)}_{n\,ij} := -\frac{1}{2k_n}(Q^{(1)}_{ni|j} + Q^{(1)}_{nj|i}), \tag{25}$$

and these modes satisfy the equation

$$Q^{(1)|i}_{n\,ij} - (k_n^2 - 2\mathbf{k}) Q^{(1)}_{nj} = 0. \tag{26}$$

The spherical harmonics are labeled by the parameter $n \in 0, \mathbf{Z}^+$ ($\vec{k} \in \mathbf{R}^3$) in the case where $\bar{\Sigma}$ is closed (open). It is useful to define the hypersurface integration operation for scalars $q(x)$ by

$$\langle\langle q \rangle\rangle := \langle q \ (g^{(3)B}/g^{(3)})^{1/2} \rangle, \tag{27}$$

which differs from the spatial average (20) by the volume element which is evaluated in the background. As we show in appendix 2.B, the spatial average (20) of a physical quantity is invariant under spatial gauge transformations, while the hypersurface integral (27) is generally gauge dependent at second and higher order in the expansion parameter of the gauge transformation.

The spherical harmonics $Q^{(0)}_n$, $Q^{(1)}_{n\,i}$, and $Q^{(2)}_{n\,ij}$ are orthogonal with respect to the hypersurface integration operation, i.e.,

$$\langle\langle Q^{(0)}_n Q^{(0)}_{n'} \rangle\rangle = \langle\langle Q^{(1)}_{n\,i} Q^{(1)i}_{n'} \rangle\rangle = \langle\langle Q^{(2)}_{n\,ij} Q^{(2)ij}_{n'} \rangle\rangle = \delta_{nn'}, \tag{28}$$

and

$$\frac{3}{2}\langle\langle Q^{(0)}_{nij}Q^{(0)ij}_{n'}\rangle\rangle = 2\langle\langle Q^{(1)}_{n\,ij}Q^{(1)ij}_{n'}\rangle\rangle = \delta_{nn'}. \tag{29}$$

Notice that the spherical harmonics are only to zeroth order orthogonal with respect to the spatial averaging operation (20) due to a generally non-vanishing first order term which arises from the expansion of the volume element $\sqrt{g^{(3)}} = \sqrt{g^B}[1 + h + O(h^2)]$.

The most general representation of a symmetric 4×4 tensor $h_{\mu\nu}$ in terms of the complete basis of spherical harmonics is given by

$$\begin{aligned} h_{00} &= -2a^2 \sum_n A_n Q^{(0)}_n, \\ h_{0i} &= -a^2 \sum_n [B^{(0)}_n Q^{(0)}_{n\,i} + B^{(1)}_n Q^{(1)}_{n\,i}], \\ h_{ij} &= 2a^2 \sum_n [H^{(0)}_{Ln} g^B_{ij} Q^{(0)}_n + H^{(0)}_{Tn} Q^{(0)}_{nij} + H^{(1)}_{Tn} Q^{(1)}_{nij} + H^{(2)}_{Tn} Q^{(2)}_{nij}], \end{aligned} \tag{30}$$

where the coefficients $A_n, B^{(0)}_n, B^{(1)}_n, H^{(0)}_{Tn}, H^{(1)}_{Tn}$, and $H^{(2)}_{Tn}$ are generally dependent on the conformal time parameter $\bar{t}$. Let u^μ be the four-velocity associated with the frame in which the energy flux of the matter vanishes, then the three-velocity u^i/u^0 associated with u^μ, can be expanded as

$$u^i/u^0 = \sum_n [v^{(0)}_n Q^{(0)}_{n\,i} + v^{(1)}_n Q^{(1)}_{n\,i}], \tag{31}$$

where $Q^{(0)}_{n\,i} := -k_n^{-1} Q^{(0)}_{n|i}$, and $u^0 = 1/a(\bar{t})$ to first order, due to the normalization $u_\mu u^\mu = -1$.

A gauge transformation is defined as a change in the correspondence between points p in S, and points $\bar{p}$ in $\bar{S}$. The most general first order gauge transformation is the result of the coordinate transformation

$$\tilde{t} = t + \sum_n T_n Q^{(0)}_n(x^\mu), \tag{32}$$

$$\tilde{x}^i = x^i + \sum_n [L^{(0)}_n Q^{(0)i}_n(x^\mu) + L^{(1)}_n\, Q^{(1)i}_n(x^\mu)], \tag{33}$$

in S, while the coordinates in $\bar{S}$ are fixed, and the correspondence between points with the same coordinates in S and $\bar{S}$ kept fixed. The coefficients T_n and L_n in expression (32) and (33) are arbitrary functions of the conformal time coordinate $\bar{t}$. Notice that T_n generates a change in the correspondence of the time coordinates in S and $\bar{S}$, while L_n generates a change in the correspondence between the spatial hypersurface coordinates on Σ and $\bar{\Sigma}$. The changes in the amplitudes of the metric tensor are calculated in the case of scalar perturbations [6],

$$\tilde{A}_n = A_n - \dot{T}_n - \frac{\dot{a}}{a} T_n, \tag{34}$$

$$\tilde{B}_n^{(0)} = B_n^{(0)} + \dot{L}_n^{(0)} + k_n T_n, \tag{35}$$

$$\tilde{H}_{Ln}^{(0)} = H_{Ln}^{(0)} - (k_n/3)L_n^{(0)} - \frac{\dot{a}}{a}T_n, \tag{36}$$

$$\tilde{H}_{Tn}^{(0)} = H_{Tn}^{(0)} + k_n L_n^{(0)}, \tag{37}$$

where an overdot denotes conformal time differentiation. For vector perturbations we find

$$\tilde{B}_n^{(1)} = B_n^{(1)} + \dot{L}_n^{(1)} \quad , \quad \tilde{H}_{Tn}^{(1)} = H_{Tn}^{(1)} + k_n L_n^{(1)}, \tag{38}$$

while for tensor perturbations the trivial relation $\tilde{H}_{Tn}^{(2)} = H_{Tn}^{(2)}$ holds for all n. The matter velocity perturbation coefficients $v_n^{(0)}$ and $v_n^{(1)}$, with respect to the coordinate frame, transform as

$$\tilde{v}_n^{(i)} = v_n^{(i)} + \dot{L}_n^{(i)}, \tag{39}$$

where $i \in \{1, 2\}$. Apart from the gauge freedom which is related to the mapping between points in S and $\bar{S}$, there is a gauge freedom related to the choice of the background scale factor $a(\bar{t})$. A first-order change in the choice of the background scale factor

$$\tilde{a}(\bar{t}) = a(\bar{t}) + D(\bar{t}) \tag{40}$$

affects the spatially homogeneous mode of the trace part of the spatial metric by a change

$$\tilde{H}_{L0}^{(0)} = H_{L0}^{(0)} + \frac{D}{a}, \tag{41}$$

where $H_{L0}^{(0)}$ is the coefficient which multiplies the spatially homogeneous trace mode in the expansion of the metric (30), and

$$H_{L0}^{(0)} = \frac{1}{3}\langle\langle \frac{\sqrt{g^{(3)}}}{\sqrt{g^B}} - 1 \rangle\rangle, \tag{42}$$

to first order.

Our approach to averaging will be based on a specification of the temporal part of the gauge (*i.e.*, the correspondence between the time coordinates t in S and $\bar{t}$ in $\bar{S}$), while maintaining covariance with respect to spatial gauge transformations (*i.e.*, the correspondence between the spatial coordinates x^i in S and $\bar{x}^i$ in $\bar{S}$).

We stress that a fully gauge covariant approach is preferred to an approach which is based on a (partially) fixed gauge, since explicitly gauge dependent results generally point out non-physical features of the calculation, including calculational mistakes. However, the intricateness of a fully gauge covariant calculation at second order makes such a calculation cumbersome (see, *e.g.*, [10]), and we will therefore follow an approach where the temporal part of the gauge is fixed, while maintaining spatial gauge covariance.

The temporal inhomogeneous part of the gauge is specified by imposing conditions on the coefficients in the expansion of the metric (30). The extrinsic curvature tensor of the constant-t hypersurfaces in S is given by,

$$K^i_j = \frac{1}{a}\sum_n \left[\frac{\dot{a}}{a} + \left(\dot{H}^{(0)}_{Ln} - \frac{\dot{a}}{a}A_n + \frac{k}{3}B^{(0)}_n\right) Q^{(0)}_n\right] \delta^i_j$$

$$+ \left[\dot{H}^{(0)}_{Tn} - kB^{(0)}_n\right] Q^{(0)i}_{n\,j} + \left[\dot{H}^{(1)}_{Tn} - kB^{(1)}_n\right] Q^{(1)i}_{n\,j} + \dot{H}^{(2)}_{Tn} Q^{(2)i}_{n\,j}. \tag{43}$$

By requiring that the coefficients A_n, $B^{(0)}_n$, and $H^{(0)}_{Ln}$ in the expansion of the metric (30) satisfy the condition

$$\dot{H}^{(0)}_{Ln} - \frac{\dot{a}}{a}A_n + \frac{k}{3}B^{(0)}_n = 0, \tag{44}$$

for all n, we specify a gauge in which the hypersurfaces of constant time t in S have spatially constant volume expansion $K = 3\dot{a}/a^2$, as is clear by contracting expression (43). Condition (44) specifies more or less uniquely a collection of spatial hypersurfaces in S (see [11]), but uniqueness is not required in the calculation which follows, since, as we will show in the following, our result for the average expansion of an inhomogeneous universe does not in relevant order depend on the choice of the inhomogeneous temporal part of the gauge.

Note that condition (44) does not constrain the choice of the time coordinate in S, and the correspondence between the time coordinates t in S and $\bar{t}$ in $\bar{S}$. We specify the time parameter t in S, up to the freedom of adding a constant, by imposing the requirement that the homogeneous component of A vanishes, *i.e.*,

$$A_0 = 0, \tag{45}$$

for all times $\bar{t}$. The choice of gauge (45) implies that the background time interval coincides with the averaged proper time interval in S, as measured by observers which are comoving with the spatial coordinates.

The gauge condition (45) can always be satisfied by performing a first order homogeneous gauge transformation. In order to clarify this statement, let us consider how equation (45) is affected by a homogeneous temporal gauge transformation. According to expression (34), a homogeneous temporal gauge transformation with $T = T_0$, induces a first order change in the metric perturbation coefficient A_0,

$$\tilde{A}_0 = A_0 - \dot{T}_0 - \frac{\dot{a}}{a}T_0. \tag{46}$$

The gauge condition (45) is satisfied by performing a gauge transformation of the form (34), where

$$T_0 = \frac{c}{a(\bar{t})} + \frac{1}{a(\bar{t})}\int^{\bar{t}} d\tau\; a(\tau)A_0(\tau), \tag{47}$$

and c is a constant of integration. The gauge condition (45) therefore determines the homogeneous temporal part of the gauge, up to a constant of integration c. According to the transformation law (36), the constant of integration c in expression (47) affects the spatially homogeneous trace part of the spatial metric. This gauge freedom can be fixed by requiring that the homogeneous trace perturbation of the spatial metric vanishes, *i.e.*,

$$H_{L0}^{(0)} = 0, \tag{48}$$

for one time $\bar{t}_c$ and for a fixed choice of the background scale factor $a(\bar{t})$ at $\bar{t} = \bar{t}_c$.

Although we have completely specified the homogeneous temporal part of the gauge by imposing the gauge conditions (45) and (48), the homogeneous trace perturbation of the spatial metric may still differ from zero for times $\bar{t} \neq \bar{t}_c$. These perturbations are related to the freedom of choosing the background scale factor $a(\bar{t})$ for times $\bar{t} \neq \bar{t}_c$, as is clear from equation (40). When we require that condition (48) holds at *all* times $\bar{t}$, then it follows from equation (40) and (41) that the choice of the background scale factor $a(\bar{t})$ is fixed for all times $\bar{t} \in \mathbf{R}$.

Recall that in section (2.2) we derived the generic linearized averaging operation for which unperturbed FLRW spacetime is a stable fixed point. It was shown that this linearized averaging operation which works on the ten components of the metric tensor, reduces to evaluating the spatial average of $\delta g^i_{\ i}$ and δg_{00}. By imposing the gauge conditions (45) and (48), we specified a choice of background geometry by requiring that the spatial averages of $\delta g^i_{\ i}$ and δg_{00} both vanish. For this choice of gauge, the averaged spacetime equals the background spacetime, and the averaging problem reduces to solving the averaged constraint equations for the background scale factor $a(\bar{t})$.

An explicit expression for the background scale factor $a(\bar{t})$ in the gauge fixed by condition (48) is obtained by substituting the expression for the background metric (21) and expression (30) for the perturbed metric, into expression (42). To first order we find

$$a^3(\bar{t}) = \langle\langle \frac{\sqrt{g^{(3)}}}{\sqrt{\eta}} \rangle\rangle, \tag{49}$$

where $g^{(3)} = \det(g_{ij})$ and $\eta = \det(\eta_{ij})$.

Recall that condition (44) fixes the inhomogeneous temporal part of the gauge, and the collection of spatial hypersurfaces on which the spatial average is evaluated. Since physical results must be gauge invariant to relevant order, one may question whether the freedom of choosing a family of hypersurfaces affects the result for the scale factor (49). It follows from the orthonormality relation (28) and the transformation property (36) that the background scale factor (49) is invariant to first order under inhomogeneous temporal gauge transformations. However, at order ϵ^2 inhomogeneous metric perturbations do contribute to the background scale factor (49), and the gauge invariance of the scale factor $a(\bar{t})$ therefore breaks down at order ϵ^2. Consistent with this limitation we will neglect terms of order ϵ^2

in our calculation, while retaining terms of order ϵ and ϵ^2/κ^2. Summarizing the content of this subsection, we completely specified the temporal and the spatially homogeneous part of the gauge, and the choice of the background, by imposing the gauge conditions (44), (45), and (48) on the metric coefficients A_n, $B_n^{(0)}$, and $H_{Ln}^{(0)}$.

2.4 Averaging the constraint equations

The classical constraint equations on a hypersurface Σ are given by

$$R^{(3)} + K^2 - K_{ij}K^{ij} = 16\pi G\rho + 2\Lambda, \tag{50}$$

$$K^j_{i;j} - K_{;i} = 8\pi G J_i, \tag{51}$$

where a semicolon denotes the covariant derivative with respect to g_{ij}, $R^{(3)}$ is the Ricci scalar associated with the induced metric g_{ij}, and

$$\rho = T^{\mu\nu}n_\mu n_\nu \ , \quad J_i = \ -T^{\mu\nu}h_{i\mu}n_\nu, \tag{52}$$

where n_μ denotes the future directed unit vector normal to Σ, and $h_{\mu\nu} := g_{\mu\nu} + n_\mu n_\nu$.

In the constant-K gauge, defined by conditions (44), (45), and (48), the constraint equation (50) takes the form

$$\frac{\dot{a}^2}{a^4} = \frac{8\pi}{3}G\rho - \frac{1}{6}R^{(3)} + \frac{1}{6}\hat{K}_{ij}\hat{K}^{ij} + \frac{1}{3}\Lambda, \tag{53}$$

where $\hat{K}^{ij} := K^{ij} - \frac{1}{3}g^{ij}K$ is the traceless part of the extrinsic curvature tensor.

In principle one could solve the constraint equation (53) for the time dependence of the scale factor $a(\bar{t})$, while taking into account all linear and higher order contributions to the right-hand side of equation (53). However, this approach is unnecessarily complicated, since all terms which do not have constant values on Σ must cancel on the right-hand side of equation (53), since the left-hand side of equation (53) is constant on Σ. For the sake of *calculational* convenience, we will take the spatial average at the right-hand side of the constraint equation (53), without changing any physical aspects of the constraint equation:

$$\frac{\dot{a}^2}{a^4} = \frac{8\pi}{3}G\langle\rho\rangle - \frac{1}{6}\langle R^{(3)}\rangle + \frac{1}{6}\langle\hat{K}_{ij}\hat{K}^{ij}\rangle + \frac{1}{3}\Lambda. \tag{54}$$

In order to solve equation (54) for the scale factor $a(t)$, we need to evaluate the spatial average of the three-curvature $R^{(3)}$, the energy density ρ, and the square of the traceless part of the extrinsic curvature tensor $\hat{K}_{ij}\hat{K}^{ij}$. We will calculate these quantities in the following three subsections.

2.5 The averaged spatial curvature

The spatial curvature perturbation $\delta R^{(3)}$ can be expanded in terms of the three-metric perturbation h_{ij} (see,*e.g.*, [9]),

$$R^{(3)} = \frac{6\mathbf{k}}{a^2} + \delta R^{(3)}, \tag{55}$$

where

$$\delta R^{(3)} = \delta R + \ \delta^2 R + O(h^3), \tag{56}$$

and

$$\delta R = h^{|k}_{|k} - h^{k|i}_{i|k}, \tag{57}$$

$$\delta^2 R = -\frac{1}{4} h^{ij} h_{ij|q}^{\ \ \ |q} + \frac{1}{2} h^{ij} h_{qi|j}^{\ \ \ |q} - \frac{1}{4} h h^{|l}_{|l} + td, \tag{58}$$

where td stands for terms which are total derivatives. Let us now evaluate the contributions to the averaged curvature perturbation $\langle R^{(3)} \rangle$ for scalar, vector, and tensor modes in the expansion of h_{ij}.

2.5.1 Scalar perturbations

It follows from expression (56) that the lowest order contribution to the spatial curvature perturbation is given by $g^{Bij}\delta R_{ij}$, which is order ϵ/κ^2, but the spatial average of this contribution vanishes to order ϵ/κ^2, due to the orthogonality relations (28). The linear curvature perturbation $g^{Bij}\delta R_{ij}$ does, however, contribute to the averaged three-curvature perturbation by a term of order ϵ^2/κ^2, *i.e.*,

$$\langle \delta R \rangle \ = \frac{12}{a^2} \sum_n (k_n^2 - 3\mathbf{k}) H^{(0)}_{Ln} (H^{(0)}_{Ln} + \frac{1}{3}\ H^{(0)}_{Tn}) \ + O(\epsilon^3/\kappa^2), \tag{59}$$

where we made use of the expansion of the volume element $\sqrt{g^{(3)}} = \sqrt{g^B}[1 + h + O(h^2)]$, and the definition of the spatial average (20).

The quadratic term $\delta^2 R$ in the expansion of the three-curvature perturbation (56) contributes to the averaged 3-curvature perturbation by a term

$$\langle \delta^2 R \rangle^{(0)}_{\infty} \ = -\frac{1}{a^2} \sum_n (k_n^2 - 3\mathbf{k}) \ (10 H^{(0)\,2}_{Ln} - \frac{2}{9} H^{(0)\,2}_{Tn} \ + \frac{8}{3} H^{(0)}_{Tn} H^{(0)}_{Ln}) + O(\epsilon^3/\kappa^2), \tag{60}$$

where we used the computer algebra package 'Maple' to derive this expression. Combining expressions (59) and (60), we find an expression for the scalar contribution to the spatial curvature perturbation,

$$\langle \delta R^{(3)} \rangle^{(0)} = \frac{2}{a^2} \sum_n (k_n^2 - 3\mathbf{k}) \phi^2_{hn} + O(\epsilon^3/\kappa^2), \tag{61}$$

where

$$\phi_{hn} := H^{(0)}_{Ln} + \frac{1}{3} H^{(0)}_{Tn} \tag{62}$$

is the gauge invariant amplitude which measures the distortion of the intrinsic geometry of the constant-K hypersurfaces. Using the expansion (56) for the spatial curvature perturbation, and the definition (62) of ϕ_{hn}, one finds that ϕ_{hn} is related to the *first order* spatial curvature perturbation by

$$\delta R = \frac{4}{a^2}\sum_n (k_n^2 - 3\mathbf{k})\phi_{hn} Q_n^{(0)}. \tag{63}$$

By substituting expression (63) into the constraint equation (50), we obtain a simple expression for ϕ_{hn} in terms of the first order energy perturbation

$$\phi_{hn} = \frac{4\pi a^2}{(k_n^2 - 3\mathbf{k})} G\bar{\rho}\, \epsilon_{hn} + O(\epsilon^2/\kappa^2) \tag{64}$$

for all n, where ϵ_{hn} is defined as the density contrast in the constant-K gauge

$$\epsilon_{hn} = \delta\rho(k_n)/\bar{\rho}, \tag{65}$$

and $\bar{\rho}$ denotes the background energy density.

We would like to express the scalar contribution to the averaged curvature perturbation in terms of observable quantities. Since the averaged curvature perturbation (61) is quadratic in ϕ_{hn}, we may use the constraint equation (64) to *first order* to determine ϕ_{hn} in terms of the fractional energy perturbation ϵ_{hn}. We obtain

$$\langle \delta R^{(3)} \rangle^{(0)} = 32(\pi a G\bar{\rho})^2 \sum_n \frac{\epsilon_h^2(q_n)}{(k_n^2 - 3\mathbf{k})} + O(\epsilon^3/\kappa^2), \tag{66}$$

where the sum (or integral when Σ is open) is taken over all possible values n. Expression (66) takes an especially simple form when expressed in terms of the power spectrum $P(k)$, which allows the representation

$$P_h(k) = \sum_n \frac{\epsilon_h^2(q_n)}{4\pi q_n^2}\, \delta(k - |q_n|), \tag{67}$$

where the subscript h refers to the constant-K gauge (see,*e.g.* [12] or [13] for more on power spectra). Combining expressions (66) and (67) yields

$$\langle \delta R^{(3)} \rangle^{(0)} = 32(\pi^2 G\bar{\rho})^2\, J_2 + O(\epsilon^2, \epsilon^3/\kappa^2), \tag{68}$$

where

$$J_2 := 4\pi a^2 \int_0^\infty dk P_h(k) \tag{69}$$

is an observable quantity known as the second moment of the power spectrum, and by absorbing a factor a^2 in the definition of J_2 we restored physical units of length.

2.5.2 Vector perturbations

Using the definition of the vector harmonics (26), and the orthogonality relations (28), we find that vector perturbations do not contribute to the spatial curvature perturbation (56). This result may be expected, since it follows from expression (38) that one can always choose a gauge in which there are no vector perturbations of the spatial metric, and the vector contribution to the averaged spatial curvature perturbation (56) must therefore vanish in any gauge, due to gauge invariance of the averaged spatial curvature perturbation.

2.5.3 Tensor perturbations

Using equations (28), and expression (56) for the second order expansion of the spatial curvature, it follows immediately that

$$\langle \delta R^{(3)} \rangle^{(2)} = \frac{1}{a^2} \sum_n k_n^2 \, H_{Tn}^{(2)\,2}, \tag{70}$$

while the tensor contribution to the term $\langle \hat{K}_{ij} \hat{K}^{ij} \rangle$ in the averaged constraint equation (53) follows immediately from the expression for the extrinsic curvature tensor (43). Although the tensor contribution to the averaged constraint equations is easily calculated in terms of the coefficients $H_{Tn}^{(2)}$, the magnitude of this term has not yet been determined quantitatively by the observation of gravitational waves.

2.6 Averaged energy density

In this subsection we will calculate the averaged energy density $\langle \rho \rangle$. In order to calculate the lowest order nontrivial contribution to the averaged energy perturbation, we will adopt the assumption in this subsection that the matter in the universe at late times after decoupling can be effectively described by the energy momentum tensor density for a pressureless perfect fluid, i.e.,

$$T^{\mu\nu} = \rho_0 u^\mu u^\nu, \tag{71}$$

where u^μ is the four-velocity of the fluid, and ρ_0 is the energy density in the rest-frame of the fluid. The equations of motion for the fluid read

$$\nabla_\mu T^{\mu\nu} = 0, \tag{72}$$

which implies

$$\partial_\mu \left(\sqrt{-g} \rho_0 u^\mu \right) = 0, \tag{73}$$

where we used that $\nabla_\mu = (\sqrt{-g})^{-1/2} \partial_\mu \sqrt{-g}$, and $u^\nu \nabla_\nu u^\mu = 0$ for a pressureless fluid. By using the spatial gauge freedom (33) we may set $B_n^{(0)} = B_n^{(1)} = 0$, such that $\sqrt{-g} = \sqrt{-g_{00}} \sqrt{g^{(3)}}$, and the equation of motion (73) takes the form

$$\partial_\mu \left(\sqrt{-g_{00}} \sqrt{g^{(3)}} \rho_0 u^\mu \right) = 0, \tag{74}$$

while in this gauge u^i/u^0 equals the matter three-velocity with respect to the normals to the constant-K hypersurfaces. The velocity four-vector u^μ can be written in the form

$$u^\mu = \left[\frac{1+v_h^2}{-g_{00}}\right]^{1/2} \delta_0^\mu + u^i \delta_i^\mu, \tag{75}$$

where $v_h^2 := g_{ij}u^i u^j$ equals to first order the square of the velocity three-vector u^i/u^0, and we used that $u^\mu u_\mu = -1$. By substituting expression (75) into the equation of motion (74), we find

$$\frac{\partial}{\partial t}\left[(1+\frac{1}{2}v_h^2)\sqrt{g^{(3)}}\rho_0\right] + \frac{\partial}{\partial x^i}\left[\sqrt{-g_{00}}\sqrt{g^{(3)}}\rho_0 u^i\right] = 0 \tag{76}$$

to first order. Using equation (76) and the definition of the spatial average (20), we obtain

$$\begin{aligned}
&\lim_{\ell\to\infty}\frac{\partial}{\partial t}\left\langle\left(1+\frac{1}{2}v_h^2\right)\rho_0\right\rangle(\ell) \\
&= -\lim_{\ell\to\infty}\left(\frac{\partial}{\partial t}\ln\tilde{N}(x,\ell)\right)\left\langle\left(1+\frac{1}{2}v_h^2\right)\rho_0\right\rangle(\ell) \\
&-\lim_{\ell\to\infty}N^{-1}(\ell)\int_{\bar{\Sigma}}dx'\,\frac{\partial}{\partial x^i}\sqrt{-g_{00}}\sqrt{g^{(3)}}\rho_0 u^i\,\theta[\ell-\Delta s(x,x')],
\end{aligned} \tag{77}$$

where $\tilde{N}(\ell)$ denotes the dimensionless quotient of $N(\ell)$, and a constant with the dimension of a three-volume. The second term on the right-hand side of equation (77) vanishes due to Gauß's theorem. Combining the remaining terms in equation (77) yields

$$\lim_{\ell\to\infty}\frac{\partial}{\partial t}\ln\left\langle\left(1+\frac{1}{2}v_h^2\right)\rho_0\right\rangle(\ell) = -\lim_{\ell\to\infty}\frac{\partial}{\partial t}\ln\tilde{N}(x,\ell). \tag{78}$$

By integrating equation (78), it follows that

$$\left\langle\left(1+\frac{1}{2}v_h^2\right)\rho_0\right\rangle(t) \propto \frac{1}{\tilde{N}(x,\ell)} \propto \frac{a^3(t_0)}{a^3(t)}, \tag{79}$$

where we used the gauge condition (48). Formula (79) shows that the rest-frame energy density ρ_0, when integrated over a spatial volume element on $\Sigma(t)$ which is comoving with the matter flow, is not conserved for a pressureless fluid, while $\rho_0(1+\frac{1}{2}v_h^2)$ is conserved to first order.

The spatial average of the energy perturbation $\delta\rho$ is obtained by expanding equation (71) for T^{00} to first order, where we use equation (75) and the gauge condition (45). We find

$$\langle\rho\rangle(t) = \langle(1+v_h^2)\rho_0\rangle(t), \tag{80}$$

which combines with equation (79),

$$\langle\rho\rangle(t) = \bar{\rho}(t) + \langle\frac{1}{2}\bar{\rho}v_h^2\rangle(t), \tag{81}$$

to first order, where we used that $\langle\rho_0\rangle(t_0)$ equals $\bar{\rho}(t_0)$ when perturbations vanish at time t_0. Indeed, the lowest order contribution to the averaged energy density (80), is given by the sum of the averaged restmass of the fluid, and the (nonrelativistic) kinetic energy of the fluid. Since v_h^2 is of the order of ϵ, the lowest order correction to the averaged energy perturbation is typically small in the observed universe, but nevertheless significant in the sense of the ambiguity which is related to the freedom of choosing a gauge and an averaging operation (see section 2.3).

It is interesting to note that there exists a simple relation due to Irvine and Layzer (see *e.g.*, [12]) which relates $W := 2\pi G\bar{\rho}J_2$, where J_2 is defined by equation (69), and the energy due to the peculiar velocity $L := \langle\frac{1}{2}\bar{\rho}v_h^2\rangle$. For a pressureless fluid and nonrelativistic motions, it can be shown that $\partial/\partial t(aW - aL) = L\dot{a}$, which, assuming that the universe departs from small values of J_2 and L and relaxes to a nearly time independent bound state at late times, implies the Newtonian virial theorem $L = W/2$.

Note that our result differs from a result derived by Futamase (see [1] and [2]), where one finds a peculiar velocity contribution to the averaged energy density which is exactly twice as large as our result which is given by equation (80). This result seems to be based on the erroneous assumption that the integral of rest-frame energy density over a spacelike hypersurface is time independent (this is only true in a gauge where $v^i - B^i$ vanishes). In this case, equation (71) yields an averaged energy perturbation which is twice the result given by equation (80). However, this result violates continuity of the scale factor at the right-hand side of equation (54) when restmass is instantaneously and homogeneously converted into kinetic energy or vice versa.

2.7 The squared shear contribution

In this subsection we will evaluate the contribution of the term

$$\langle\hat{K}_{ij}\hat{K}^{ij}\rangle, \tag{82}$$

in the averaged constraint equations (54), for scalar and vector perturbations. The scalar and vector part of $\hat{K}_{ij}$ are coupled to the matter current by the constraint equation (51), which takes the form

$$\hat{K}^j_{i;j} = 8\pi G J_i, \tag{83}$$

when evaluated in the constant-K gauge. The matter current J_i is defined by expression (52), and can be expanded as

$$J_i = (\bar{\rho} + \bar{P}) \sum_n \left[v^{(0)}_{hn} Q^{(0)}_{ni} \; + v^{(1)}_{hn} Q^{(1)}_{ni} \right], \tag{84}$$

to first order, where $v^{(0)}_{hn}$ and $v^{(1)}_{hn}$ denote the scalar and vector components of the velocity three-vector of the matter with respect to the normals to the constant-K

hypersurfaces, $Q^{(0)}_{ni} := -k^{-1}Q^{(0)}_{n|i}$, and $\bar{\rho}$ and $\bar{P}$ denote the background energy and pressure density.

By substituting the traceless part of the extrinsic curvature tensor (43), and the expansion (84) for J_i, into the constraint equation (83), we obtain

$$\frac{2}{3}(k_n - 3\mathbf{k}/k_n)[\dot{H}^{(0)}_{Tn} - k_n B^{(0)}_n] = \; aG(\bar{\rho} + \bar{P})\; v^{(0)}_{hn} \tag{85}$$

for scalar perturbations, and

$$\frac{1}{2}(k_n - 2\mathbf{k}/k_n)[\dot{H}^{(1)}_{Tn} - k_n B^{(1)}_n] = \; aG(\bar{\rho} + \bar{P})\; v^{(1)}_{hn} \tag{86}$$

for vector perturbations. Expressions (85) and (86) yield expressions for the scalar and vector traceless part of the extrinsic curvature tensor (43), in terms of the matter velocity, which can be used to evaluate the scalar and vector contribution to expression (82). For scalar perturbations we find,

$$\langle \hat{K}^{(0)}_{ij} \hat{K}^{(0)ij} \rangle \; = \frac{3}{2} a^2 G^2 (\bar{\rho} + \bar{P})^2 \; \sum_n \frac{v^{(0)2}_{hn}}{(k_n - 3\mathbf{k}/k_n)^2}, \tag{87}$$

and for vector perturbations

$$\langle \hat{K}^{(1)}_{ij} \hat{K}^{(1)ij} \rangle \; = 2a^2 G^2 (\bar{\rho} + \bar{P})^2 \; \sum_n \frac{v^{(1)2}_{hn}}{(k_n - 2\mathbf{k}/k_n)^2}. \tag{88}$$

The coupling between the matter current and the shear of the normals to the constant-K hypersurfaces, can be interpreted as the 'frame dragging' effect which occurs in the presence of moving matter (e.g., as in the region around a rotating black hole). It follows from expressions (87) and (88), taking into account the normalizations of the scalar and vector modes (see expression (29)), that the matter current and $\hat{K}_{ij}$ couple with different strength for scalar and vector perturbations. Furthermore, the strength of the coupling vanishes proportional to k_n^{-1} when $k_n \rightarrow \infty$. Since $v^2_{hn} = O(\epsilon)$, when velocity perturbations are generated by density perturbations at late times, it follows that expressions (87) and (88) contribute to the averaged constraint equations (54) by a term of order $\epsilon\kappa^2$, which is negligible compared to the leading order kinetic energy contribution discussed in section (2.6) for perturbations at length scales much smaller than the Hubble radius.

However, for perturbations at arbitrary large length scales, the strength of the coupling grows proportional to δ^{-1} when $\delta \downarrow 0$, where $\delta := k_n^2 - 3\mathbf{k}$ for scalar perturbations and $\delta := k_n^2 - 2\mathbf{k}$ for vector perturbations. Note that since k_n must be real for bounded solutions, the limit $\delta \downarrow 0$ does not exist when $\mathbf{k} < 0$, and the limit $\delta \downarrow 0$ does not exist when $\mathbf{k} > 0$ since k_n takes only discrete values in this case. Note that the divergent coupling between the metric and the matter velocity

for $\delta \downarrow 0$ and $\mathbf{k} = 0$, is unrelated to the dynamics of the matter and metric at small scales and late times, since perturbations for which $\delta \ll 1$ are typically larger than the Hubble radius, and must have a primordial origin.

A natural question which arises is whether the divergence in equations (87) and (88) for $\delta \downarrow 0$ can be purely attributed to a large warping of the constant-K hypersurfaces, which can be removed by choosing another gauge. Indeed, it follows from expressions (43) and (35) that the scalar part of $\hat{K}^i_j$ can be set equal to zero, by a temporal gauge transformation with $T = k^{-2}[\dot{H}^{(0)}_{Tn} - kB^{(0)}_n]$, but according to expressions (63), (36) and (37), the *intrinsic* spatial curvature does diverge when $\delta \downarrow 0$ in this gauge. Furthermore, due to expression (38), the vector part of $\hat{K}^i_j$ is gauge invariant, and the divergence in equation (88) is therefore independent of the choice of timeslicing. From the point of view of the matter, the most natural choice of gauge is a comoving time-orthogonal gauge, which is defined by the condition that the spatial coordinates are comoving with the normals to the constant-t hypersurfaces (i.e., $B^{(0)} = B^{(1)} = 0$), and the scalar part of the matter velocity with respect to the normals to the constant-t hypersurfaces vanishes (i.e., $v^{(0)} - B^{(0)} = v^{(0)} = 0$). According to expressions (35) and (39), a gauge transformation from the constant-K gauge to a comoving time-orthogonal gauge is generated by $T = k^{-1}v^{(0)}_{hn}$. In this gauge, the scalar part of the shear of the *matter* coincides with the scalar part of $\hat{K}^i_j$. By transforming equation (85) from the constant-K gauge to a comoving gauge, we find that the infrared divergence of the scalar part of $\hat{K}^i_j$ has the same strength in both gauges, and its presence is therefore related to the presence of shearing matter. At first sight, a divergence of the shear of the matter for $\delta \downarrow 0$ seems to be inconsistent with the smallness of the velocity perturbations which are the source of the metric perturbations. There is no real inconsistency however, since the matter velocity perturbation is gauge dependent, and it might therefore be anomalously small in the constant-K gauge, without being in conflict with large matter shear perturbations. These observations show that the divergence in equations (87) and (88) is of a physical nature.

The absence of FLRW solutions of the constraint equations (83) when homogeneous vector perturbations of the matter velocity are present, might seem peculiar, since solutions of the Einstein equations correspond to stable points of the action. At this point we should recall that we have limited our scope to FLRW background spacetimes, which are by definition spatially homogeneous and isotropic. In the presence of homogeneous matter velocity perturbations, our spacetime is no longer isotropic in the averaged sense, and there is no FLRW background solution which is everywhere close to our perturbed spacetime. A satisfactory description of homogeneous velocity perturbations about FLRW spacetimes requires an extension of the class of background solutions which includes those models which are homogeneous but not necessarily isotropic. These solutions are given by the Bianchi models of type V and VII_h, which include FLRW with $\mathbf{k} = -1$ as a special case, and type VII_0 which includes the FLRW geometry with $\mathbf{k} = 0$ (see *e.g.*, [14] – [16]).

2.8 The averaged expansion

By substituting the expressions for the averaged curvature perturbation and the averaged energy density, which were derived in the previous subsections 2.5, 2.6 and 2.7, into the averaged constraint equation (54), we obtain,

$$\frac{\dot{a}^2}{a^4} = \frac{8\pi}{3}G\bar{\rho} - \frac{\mathbf{k}}{a^2} + \frac{1}{3}\Lambda + \frac{8\pi G}{6}\langle\bar{\rho}v_h^2\rangle - \frac{32\pi^2}{3}(G\bar{\rho})^2 J_2 \qquad (89)$$

$$+gw + O(\epsilon^2, \epsilon\kappa^2, \epsilon^3/\kappa^2),$$

where J_2 is defined by equation (69), and the term gw denotes the contribution due to gravitational waves (see subsection 2.5.3). We see that the averaged constraint equation (89) takes the form of the standard Friedmann equation, plus a contribution due to the peculiar velocity of the matter, and a contribution due to the averaging of scalar and tensor metric perturbations. Let us now determine the magnitudes of the different contributions on the right-hand side of equation (89) by means of the observational values for $\bar{\rho}$ and J_2. Estimates from the Lick and CfA catalogs [17] [18] value $J_2 \approx 200h^{-2}$ Mpc^2, and $\bar{\rho} \approx 1.88 \times 10^{-29}h^2\Omega$ g cm^{-3}, where h is a dimensionless factor which expresses the uncertainty in the value of the Hubble parameter $H_0 = 100h$ kms^{-1}Mpc^{-1}, and h is believed to be between 0.5 and 0.85. Inserting these values in the different terms on the right-hand side of equation (89), one finds,

$$\frac{8\pi G}{3}\bar{\rho} = 1.14 \times 10^{-35}h^2\Omega\mathrm{s}^{-2}, \qquad (90)$$

$$\frac{32\pi^2}{3}(G\bar{\rho})^2 J_2 \sim 1.0 \times 10^{-39}h^2\Omega^2\mathrm{s}^{-2}, \qquad (91)$$

and

$$\frac{8\pi G}{6}\langle\bar{\rho}v_h^2\rangle \sim 1.3 \times 10^{-40}h\ \Omega\mathrm{s}^{-2}, \qquad (92)$$

where we used the relation $v \sim (3\pi G\bar{\rho}J_2)^{1/2}$ (see section 2.6). According to equations (90)– (92), and the constraint equation (89), the matter induced metric inhomogeneities act as a very small negative correction to the averaged energy density, equal to about 1.0 $\times\Omega$ part in 10^4, while the backreaction due to the peculiar velocity of the matter acts as a positive correction to the averaged energy density, equal to about 1.2 parts in 10^5. The small negative correction to the averaged energy density leads to a slight overestimation of the age of the universe $t_0 = \frac{2}{3}H_0^{-1}$ assuming that $\Omega = 1$, equal to about 5 parts in 10^5.

2.9 Comparison with previous work

The work on this chapter started as a correction of the derivation by Futamase in [1] [2] on the points of the treatment of the gauge freedom (see section 2.3) and the

choice of the averaging operation (see section 2.2). This chapter was also inspired as an attempt to address the fundamental ambiguity which enters the calculation of any averaged metric through the freedom of choosing an averaging operation.

In a recent independent paper by Russ *et al.* [21], the backreaction due to density perturbations was calculated by using the relativistic Zel'dovich approximation in a comoving gauge. The expression derived by Russ *et al.* for the backreaction due to matter density perturbations agree in sign, but is roughly an order of magnitude larger than the result derived in this chapter. Furthermore, a possible effect due to vorticity of the matter was ignored in that paper. It should be noted that direct comparison between the results by Russ *et al.* and the results derived in this chapter, is nontrivial due to the fact that the gauges used in the paper by Russ *et al.* and this chapter are not related by a first-order gauge transformation. Namely, a gauge transformation from the constant-K gauge to the comoving synchronous gauge requires $\dot{L}_n = -v_{hn}$ due to equation (39), and $v_{hn} = O(\epsilon^{1/2})$ since $v_{hn}^2 = O(\epsilon)$ when velocity perturbations are generated by density perturbations at late times. By working in a constant-K gauge, we avoided the problem of a breakdown of the perturbative expansion which occurs in the comoving gauge (namely, since metric and matter density perturbations are of the same magnitude in a comoving gauge, metric perturbations get typically large at late times, even though the perturbations in the intrinsic geometry are generally small in the observed universe).

Finally, we mention the paper by Buchert and Ehlers [20], where one integrates the Raychaudhuri equation over a spatial hypersurface in a Newtonian background, and a globally vanishing correction to the averaged expansion was found. Although the Raychaudhuri equation is also valid in General Relativity, the Newtonian approximation enters the calculation where the correction to the averaged expansion is expressed in terms of a boundary term, which accounts for the difference between the Newtonian result and the nontrivial correction (90) to the averaged energy density derived in this chapter.

2.10 Conclusions

We derived the generic linearized averaging operation for metrics starting from the requirement that unperturbed FLRW spacetime is a stable fixed point of the averaging operation. By a gauge invariant approach, we eliminated unphysical degrees of freedom in our problem, and we clarified the fundamental ambiguities which are related to the freedom of fitting the averaged spacetime to the inhomogeneous spacetime. The leading order nontrivial corrections to the standard Friedmann equation are expressed in terms of the power spectrum of the matter, and the effect is calculated quantitatively by means of the observational data. The dominant correction to the averaged expansion is caused by the backreaction of matter density perturbations, and leads to a slower expansion rate and an overestimation of the age of the universe by approximately 5 parts in 10^5. The backreaction of velocity perturbations, including vortical motion of the matter, appears to be negligible

at small length scales. However, it was shown that the backreaction of velocity perturbations can be significant in the large wavelength limit.

2.A Averaging and gauge invariance

In this appendix we discuss the relation between the volume element in the hypersurface integral, and gauge invariance at second and higher order in the expansion parameter of the gauge transformation. Let ϕ be a one parameter group of diffeomorphisms $\phi : \mathbf{R} \times \Sigma \to \Sigma$, which is defined by the condition that $\phi_{\lambda=0}$ is the identity, and the curves $\phi_\lambda(p)$ are integral curves of a vector field ξ in Σ (see, *e.g.*, [19] and [10] for the mathematical details which are involved). A gauge is specified by choosing a mapping between points p in Σ, and points $\bar{p}$ in $\bar{\Sigma}$. Assuming that a choice of gauge has been made, then a one parameter *group* of gauge choices is obtained by mapping the points $\phi_\lambda(p)$ in Σ to points $\bar{p}$ in $\bar{\Sigma}$, for all $\lambda \in \mathbf{R}$ (the more generic case of a one parameter *family* of mappings of points in the background and the perturbed spacetime, is discussed in [10], but there is no need to introduce this complication in the derivation which follows).

Let us now consider a scalar function $q(x)$, which lives in Σ (such that its value in a point p in Σ is fixed, while its value in a point $\bar{p}$ in $\bar{\Sigma}$ depends on the choice of gauge). Let us recall that the spatial average and the hypersurface integral of $q(x)$, are related by

$$\langle q \rangle = \langle\langle q \ (g^{(3)}/g^B)^{1/2} \rangle\rangle, \tag{1}$$

where we used the definition (27). The integrand at the right-hand side of equation (1) is gauge dependent, and may be expanded in powers of λ about $\lambda = 0$, i.e.,

$$q(g^{(3)}/g^B)^{\frac{1}{2}}(\lambda, \bar{p}) = \sum_{k=0}^{k=\infty} \frac{\lambda^k}{k!} \mathcal{L}_\xi^k \ q(g^{(3)}/g^B)^{1/2}, \tag{2}$$

where $\mathcal{L}_\xi^k$ denotes the kth order Lie derivative with respect to ξ, evaluated in $\bar{p}$. By substituting the expansion (2) in the integrand at the right-hand side of equation (1), we obtain

$$\langle q \rangle(\lambda) - \langle q \rangle(0) = \sum_{k=1}^{k=\infty} \frac{\lambda^k}{k!} \ \langle\langle \mathcal{L}_\xi^k q (g^{(3)}/g^B)^{1/2} \rangle\rangle. \tag{3}$$

For $k = 1$, the contribution to the right-hand side of equation (3) is evaluated using

$$\mathcal{L}_\xi q = \xi^i q_{;i} \tag{4}$$

and

$$\mathcal{L}_\xi (g^{(3)}/g^B)^{1/2} = \xi^i_{;i} (g^{(3)}/g^B)^{\frac{1}{2}}, \tag{5}$$

where the semicolon denotes covariant differentiation with respect to g_{ij}. Combining equations (4) and (5) yields,

$$\mathcal{L}_\xi q (g^{(3)}/g^B)^{1/2} = \ (q\xi^i)_{;i} (g^{(3)}/g^B)^{1/2}, \tag{6}$$

and

$$\langle\langle (q\xi^i)_{;i} (g^{(3)}/g^B)^{1/2} \rangle\rangle = 0, \tag{7}$$

due to Gauss's theorem. The $k = 2$ contribution to the right-hand side of equation (3) is obtained by making the substitution $q \to (q\xi^i)_{;i}$ in expression (6), and for arbitrary $k \in \mathbf{Z}^+$ the same result follows by induction. Since the terms at the right-hand side of equation (3) vanish for all k, we established that the spatial average of a scalar function q is gauge invariant to arbitrary order in the expansion parameter λ. Applying the same analysis as above to the hypersurface integral of a scalar field $q(x)$, we find,

$$\langle\langle q\rangle\rangle(\lambda) \; - \langle\langle q\rangle\rangle(0) = \sum_{k=1}^{k=\infty} \frac{\lambda^k}{k!} \, \langle\langle \mathcal{L}_\xi^k q\rangle\rangle, \tag{8}$$

which depends on λ, due to equation (4), unless q is a constant on Σ. A similar derivation, where we reverse the roles of the spacetimes Σ and $\bar{\Sigma}$, shows that the hypersurface integral of a scalar field $\bar{q}(\bar{x})$ which lives in $\bar{\Sigma}$ is gauge invariant, while the spatial average of $\bar{q}(\bar{x})$ is gauge invariant iff $\bar{q}(\bar{x})$ is constant in $\bar{\Sigma}$. It follows from these observations that the spatial average of a perturbation $\delta q := q(x) - \bar{q}(\bar{x})$ is gauge invariant iff $\bar{q}(\bar{x})$ is constant on $\bar{\Sigma}$. Note that Futamase in [1] uses the hypersurface integral as a spatial averaging operation in the calculation of second order effects, while he does not consistently fix a gauge in these papers (namely, in this paper one assumes a comoving synchronous gauge *and* constant expansion on the hypersurfaces of constant time coordinate).

2.B Uniqueness of the averaging operation

In this appendix we derive the decomposition of the generic linearized averaging operation

$$\hat{A}^{\infty} := \lim_{n\to\infty} \hat{A}^{(1)n}\delta g_{\mu\nu}, \tag{9}$$

in terms of the spatial average $\langle \delta g_{\mu\nu}\rangle(t)$, which is uniquely defined. Note that the existence of the limit (9) implies that

$$\hat{A}^{(1)}\delta g^{*}_{\mu\nu} = \delta g^{*}_{\mu\nu} \tag{10}$$

for arbitrary spatially homogeneous and isotropic perturbations $\delta g^{*}_{\mu\nu}$ (up to the freedom of diffeomorphisms acting at either side of equation (10)).

Without loss of generality, a spatially homogeneous and isotropic perturbation $\delta g^{*}_{\mu\nu}$ about $\bar{\mathbf{S}}$ can be written in the form

$$\delta g^{*}_{\mu\nu} \ = \phi_1(\bar{t})\bar{n}_\mu\bar{n}_\nu + \phi_2(\bar{t})\bar{h}_{\mu\nu}, \tag{11}$$

where $\bar{h}_{\mu\nu} := g^{B}_{\mu\nu} + \bar{n}_\mu\bar{n}_\nu$, and $\bar{n}_\mu$ denotes the timelike future directed vector in $\bar{\mathbf{S}}$ which is orthogonal to $\bar{\Sigma}$, and which is normalized with respect to the background metric $g^{B}_{\mu\nu}$, and $\phi_1(\bar{t})$ and $\phi_2(\bar{t})$ are arbitrary functions of $\bar{t}$.

When we substitute expression (11) for $\delta g^{*}_{\mu\nu}$ and expression (15) for $\hat{A}^{(1)}$ into condition (10), we obtain,

$$\int dt' d^3x' \left[\ \phi_1(t')\bar{n}_\rho\bar{n}_\sigma \ + \phi_2(t')\bar{h}_{\rho\sigma}\right]\ f^{\rho\sigma}_{\mu\nu}(x,x')$$

$$= \phi_1(t)\bar{n}_\mu(x)\bar{n}_\nu(x) + \phi_2(t)\bar{h}_{\mu\nu}(x), \tag{12}$$

for arbitrary functions $\phi_1(t)$ and $\phi_2(t)$. Equation (12) holds for arbitrary $\phi_1(t)$ and $\phi_2(t)$ if and only if

$$\int_{\bar{\Sigma}} d^3x'\ \bar{n}_\rho(x')\bar{n}_\sigma(x')\ f^{\rho\sigma}_{\mu\nu}(x,x') \ = \delta(t'-t)\bar{n}_\mu(x)\bar{n}_\nu(x) \tag{13}$$

and

$$\int_{\bar{\Sigma}} d^3x'\ \bar{h}_{\rho\sigma}(x')\ f^{\rho\sigma}_{\mu\nu}(x,x') \ = \delta(t'-t)\bar{h}_{\mu\nu}(x). \tag{14}$$

Expression (13) shows that $f^{\rho\sigma}_{\mu\nu}(x,x')$ is proportional to a delta distribution $\delta(t-t')$. It follows from this observation that $\hat{A}^{(1)}$ can be naturally defined in terms of a linearized *spatial* averaging operation $\hat{A}^{(1)}_s$, *i.e.*,

$$\hat{A}^{(1)}\delta g_{\mu\nu} \ = \hat{A}^{(1)}_s(t)\delta g_{\mu\nu}, \tag{15}$$

where $\hat{A}^{(1)}_s$ is defined by

$$\hat{A}^{(1)}_s\delta g_{\mu\nu} = \int_{\bar{\Sigma}(t)} d^3x'\ f^{\rho\sigma}_{\mu\nu}(t,x^i,x^{i\prime})\ \delta g_{\rho\sigma}(x^{i\prime}) \tag{16}$$

and

$$f^{\rho\sigma}_{\mu\nu}(t, x^i, x^{i\prime}) := \int_{\Delta t'} dt' \; f^{\rho\sigma}_{\mu\nu}(t, t', x^i, x^{i\prime}), \tag{17}$$

and $\Delta t'$ is chosen such that $t \in \Delta t'$. At first sight, the decomposition of the linear averaging operation $\hat{A}^{(1)}$ in terms of a spatial averaging operation which is defined on a collection of spatial hypersurfaces might be surprising, since the choice of a collection of spatial hypersurfaces $\bar{\Sigma}(t)$ in $\mathbf{S}$ is gauge dependent. It was shown in section (2.3) that although the choice of $\bar{\Sigma}(t)$ in $\mathbf{S}$ is gauge dependent, the linearized spatial averaging operation (15) is to first order gauge independent.

Assuming that the limit in equation (17) exists, then by substituting expression (15) into expression (9) one finds that the limit

$$\langle \delta g_{\mu\nu} \rangle := \lim_{n\to\infty} \hat{A}^{(1)n}_s \delta g_{\mu\nu} \tag{18}$$

exists. We will show that the limiting spatial averaging operation which is defined by equation (18) is universal.

Expression (16) and the definition (18) imply that

$$\langle \delta g_{\mu\nu} \rangle := \lim_{n\to\infty} \int_{\bar{\Sigma}(t)} d^3x' \; f^{n\rho\sigma}_{\mu\nu}(t, x^i, x'^i) \; \delta g_{\rho\sigma}(x'^i), \tag{19}$$

where $f^{n\rho\sigma}_{\mu\nu}$ is defined in terms of $f^{\rho\sigma}_{\mu\nu}$ by induction over n,

$$f^{n\rho\sigma}_{\mu\nu}(x^i, x'^i) = \int_{\bar{\Sigma}(t)} d^3q f^{\rho\sigma}_{\alpha\beta}(t, x'^i, q^i) \; f^{n-1\alpha\beta}_{\mu\nu}(t, x^i, q^i), \tag{20}$$

and $f^{1\rho\sigma}_{\mu\nu} := f^{\rho\sigma}_{\mu\nu}$. Let us now try to determine the limit

$$f^{\infty\rho\sigma}_{\mu\nu} := \lim_{n\to\infty} f^{n\rho\sigma}_{\mu\nu}. \tag{21}$$

An explicit calculation of $f^{\infty\rho\sigma}_{\mu\nu}$, using the definition definition (20) for $f^{n\rho\sigma}_{\mu\nu}$ and starting with arbitrary realizations for $f^{\rho\sigma}_{\mu\nu}$, would be quite cumbersome, but fortunately it appears that the symmetries of the background spacetime $\bar{S}$, and the stability condition (17) determine $f^{\infty\rho\sigma}_{\alpha\beta}$ completely.

Recall that we required that the limit (17) converges to a spatially homogeneous and isotropic metric perturbation for arbitrary perturbations $\delta g_{\mu\nu}$, which implies that

$$\langle \delta g_{\mu\nu} \rangle (x^i) = \int_{\bar{\Sigma}(t)} d^3x' \; f^{\infty\rho\sigma}_{\mu\nu}(t, x^i, x'^i) \; \delta g_{\rho\sigma}(x'^i) = \delta g^*_{\mu\nu}, \tag{22}$$

for all x, where we used expression (15) and $\delta g^*_{\mu\nu}$ has the form (11). If expression (22) holds for arbitrary perturbations $\delta g_{\rho\sigma}(x'^i)$, it also holds for arbitrary perturbations $\delta g_{\rho\sigma}(x'^i + c^i)$, where $c^i \in \mathbf{R}$. By absorbing the constants c^i into the coordinates x^i, one finds that expression (22) remains unchanged under the substitution

$$f^{\infty\rho\sigma}_{\mu\nu}(x^i, x'^i) \to \; f^{\infty\rho\sigma}_{\mu\nu}(x^i, x'^i - c^i). \tag{23}$$

Furthermore, since the right-hand side of equation (22) is spatially homogeneous by requirement, we find that the left-hand side of equation (22) must be also invariant under the substitution

$$f^{\infty\rho\sigma}_{\mu\nu}(x^i, x'^i) \rightarrow \; f^{\infty\rho\sigma}_{\mu\nu}(x^i + d^i, x'^i), \tag{24}$$

where $d^i \in \mathbf{R}$ is arbitrary. Since equation (22) is invariant under the operations (23) and (24) for arbitrary perturbations $\delta g_{\mu\nu}$, we conclude that $f^{\infty\rho\sigma}_{\mu\nu}$ is (up to the freedom of performing diffeomorphisms) constant on Σ when regarded as a distribution (*i.e.*, neglecting sets of Lebesque measure zero). Furthermore, since equation (22) holds for arbitrary $\delta g_{\mu\nu}$, the distribution $f^{\infty\rho\sigma}_{\mu\nu}(x, x')$ must be proportional to a tensor of the form (11) in the point x, thereby fixing the $\mu\nu$ dependent part of $f^{\infty\rho\sigma}_{\mu\nu}$. We may therefore write

$$f^{\infty\rho\sigma}_{\mu\nu}(x, x') \; = g_1^{\rho\sigma}(x')\bar{n}_\mu(x)\bar{n}_\nu(x) \; + g_2^{\rho\sigma}(x')\bar{h}_{\mu\nu}(x), \tag{25}$$

where $g_1^{\rho\sigma}(x')$ and $g_2^{\rho\sigma}(x')$ are spatially homogeneous tensor densities in x', and we used expression (11) for $g^*_{\mu\nu}$.

A similar argument, using the invariance of equation (22) under the group of spatial rotations (using that $\mathcal{F}$ does not explicitly depend on x), shows that the bi-tensor density $f^{\infty\rho\sigma}_{\mu\nu}(x, x')$ is isotropic with respect to the indices σ and ρ, which implies that the tensor densities $g_1^{\rho\sigma}(x')$ and $g_2^{\rho\sigma}(x')$ in expression (25) are of the form,

$$g_1^{\rho\sigma}(x') = \alpha_1\sqrt{g^{(3)}}\; \bar{n}^\rho(x')\bar{n}^\sigma(x') \; + \alpha_2\sqrt{g^{(3)}}\bar{h}^{\rho\sigma}(x') \tag{26}$$

and

$$g_2^{\rho\sigma}(x') = \alpha_3\sqrt{g^{(3)}}\; \bar{n}^\rho(x')\bar{n}^\sigma(x') \; + \alpha_4\sqrt{g^{(3)}}\bar{h}^{\rho\sigma}(x'), \tag{27}$$

where $g^{(3)}$ denotes the real space volume element, which follows from requiring spatial gauge invariance at higher orders (see appendix 2.A), and the factors α_n ($n \in \{1, 2, 3, 4\}$) are constant on Σ. Substituting expressions (26) and (27) in expression (25) yields

$$f^{\infty\rho\sigma}_{\mu\nu}(x, x') \; = \sqrt{g^{(3)}}[\alpha_1\bar{n}^\rho(x')\bar{n}^\sigma(x')\; \bar{n}_\mu(x)\bar{n}_\nu(x) \; + \alpha_4\bar{h}^{\rho\sigma}(x')\bar{h}_{\mu\nu}(x)], \tag{28}$$

where we used expressions (13) and (14) to show that the terms proportional to α_2 and α_3 vanish.

By substituting expression (28) for $f^{\infty\rho\sigma}_{\mu\nu}$ into condition (22), where we set $\delta g_{\rho\sigma}$ equal to $\delta g^*_{\mu\nu}$ defined by expression (11), we find that the constants α_1 and α_4 in expression (28) must satisfy the condition

$$\int_{\bar{\Sigma}(t)} d^3x'\sqrt{g^{(3)}}\alpha_1 = \; 3\int_{\bar{\Sigma}(t)} d^3x'\sqrt{g^{(3)}}\alpha_4 = 1. \tag{29}$$

Expression (29) shows that the constants α_1 and $3\alpha_4$ are equal to $(\text{volume}(\Sigma))^{-1}$ when Σ is closed, while in the case when Σ is open, α_1 and α_4 are defined in a

distributional sense by condition (29), and by the condition that α_1 and α_4 are constant on $\Sigma(t)$.

By substituting expression (28) into expression (22) we obtain the explicit expression for the spatial average,

$$\langle \delta g_{\mu\nu} \rangle = \int_{\bar{\Sigma}(t)} d^3x' \sqrt{g^{(3)}} \alpha_1 \tag{30}$$

$$\left[\bar{n}^\rho(x') \bar{n}^\sigma(x') \; \bar{n}_\mu(x) \bar{n}_\nu(x) \; + \frac{1}{3} \bar{h}^{\rho\sigma}(x') \bar{h}_{\mu\nu}(x) \right] \; \delta g_{\rho\sigma}. \tag{31}$$

Note that $\bar{n}^\rho \bar{n}^\sigma \delta g_{\rho\sigma}$ equals the perturbation of g_{00} in coordinates which are synchronous in the background (i.e., coordinates for which $g^B_{\mu 0} = -\delta^0_\mu$), while $\bar{h}^{\rho\sigma} \delta g_{\rho\sigma}$ equals the perturbation of the spatial volume element on Σ, to first order.

Summarizing the derivation in this appendix, we showed that the general linearized averaging operation which is a functional of metric perturbations about FLRW spacetime, and for which unperturbed FLRW spacetime is a stable fixed point, has a unique limit when applied iteratively to perturbed FLRW spacetime. Furthermore, we showed that this linearized averaged operation is naturally defined in terms of a spatial averaging operation which works on g_{00} and the spatial volume perturbation in coordinates which are synchronous in the background.

References

[1] T. Futamase, Phys. Rev. **D 53**, 681 (1996).

[2] S. Bildhauer and T. Futamase, Gen. Relativ. Gravit. **23**, 1251 (1991).

[3] R. M. Zalaletdinov, Gen. Relativ. Gravit. **24**, 1015 (1992).

[4] R. M. Zalaletdinov, Gen. Relativ. Gravit. **25**, 673 (1993).

[5] M. Carfora and A. Marzuoli, Phys. Rev. Lett. **53**, 2445 (1984).

[6] J. M. Bardeen, Phys. Rev. **D 22**, 1882 (1980).

[7] C. C. Dyer and R. C. Roeder, Astrophys. J. **172**, L115 (1972).

[8] G. F. R. Ellis, *General Relativity and Gravitation*, eddited by B. Bertotti *et al.* (Reidel, Dordrecht, 1984), 215.

[9] G.'t Hooft and M. Veltman, Ann. In. Henri Poincare **20**, 69 (1974).

[10] M. Bruni *et al.* Class. Quantum Grav. **14**, 2585 (1997).

[11] D. M. Eardley and L. Smarr, Phys. Rev. **D 19**, 2239 (1979).

[12] P. J. E. Peebles, *Principles of Physical Cosmology* (Princeton University Press, Princeton, 1993) 508.

[13] P. Coles and F. Lucchin, *The Origin and Evolution of Cosmic Structure* (J. Wiley, New York, 1995) 265.

[14] C. B. Collins and S. W. Hawking, Mon. Not. R. Astron. Soc. **162**, 307 (1973).

[15] C. B. Collins and G. F. R. Ellis, Phys. Rep. **56**, 65 (1979).

[16] C. G. Hewitt *et al.*, in *Dynamical Systems in Cosmology*, edited by J. Wainwright and G. F. R. Ellis (Cambridge University Press, Cambridge, 1997) chap. 9.

[17] M. Clutton-Brock and P. J. E. Peebles, Astron. J. **86**, 1115 (1981).

[18] M. Davis and P. J. E. Peebles, Astrophys. J. **267**, 465 (1983).

[19] R. M. Wald, *General Relativity* (University of Chicago Press, Chicago, 1984) 437.

[20] T. Buchert and J. Ehlers, Astron. Astrophys. **320**, 1 (1997).

[21] H. Russ *et al.*, Phys. Rev. **D 53**, 6881 (1996).

3 Variational dynamics in open spacetimes

Abstract

We study the effect of non-vanishing surface terms at spatial infinity on the dynamics of a scalar field in an open Friedmann-Lemaître-Robertson-Walker (FLRW) spacetime. Starting from the path-integral formulation of quantum field theory, we argue that classical physics is described by field configurations which extremize the action functional in the space of field configurations for which the variation of the action is well defined. Since these field configurations are not required to vanish outside a bounded domain, there can be a non-vanishing contribution of a surface term to the variation of the action. We then investigate whether this surface term has an effect on the dynamics of the action-extremizing field configurations. This question appears to be surprisingly nontrivial in the case of the open FLRW geometry since surface terms tend to grow as fast as volume terms in the infinite volume limit. We find that surface terms can be important for the dynamics of the field at a classical and quantum level, when there are supercurvature perturbations.

3.1 Introduction

The idea that surface terms can be important when the Lagrangian method is applied to cosmology has been studied earlier in the context of spatially homogeneous but anisotropic models [1] - [4]. In this case, a surface term appears when the Lagrangian is varied with respect to a spatially homogeneous metric perturbation, and the assumption of spatial homogeneity prevents the vanishing of this term when it is evaluated on an arbitrarily distant compact two-surface. In most other cases where the variational approach is applied to cosmology, surface terms are made to vanish trivially by evaluating only variations with respect to variables which vanish outside a bounded domain. The justification for this approach seems to be that one recovers the 'correct' field equations, which are the standard Euler-Lagrange equations. In a cosmological context, this way of reasoning can be questioned, both from a theoretical and an observational point of view. From observations, it is not *a priori* clear which are the correct equations of motion describing the dynamics of fields at length scales larger than the observable universe, and in different cosmological models. From a theoretical point of view, the relation between extremal action fields and classical physics has a natural foundation in quantum field theory. However, field configurations which vanish outside a bounded domain do not play a central role in quantum field theory, and this assumption may be questioned in a cosmological context when the spacetime itself is not bounded.

In this chapter we will study this situation by means of an idealized model, which consist of a Klein-Gordon field in both a spatially flat and a spatially open FLRW universe. Our motivation for studying a scalar field stems from the aim to keep our equations simple, and the possible importance of these fields in the description of the early universe. We will concentrate on the open FLRW geometry, since this spacetime has some specific properties which allow surface terms to become important.

One of these properties is that eigenfunctions of the spatial Laplacian occur in two types. First, there are eigenfunctions with eigenvalues exceeding $\frac{1}{6}$ times the spatial curvature, and these eigenfunctions are complete in the space of square integrable functions [5]. Second, there are eigenfunctions of the spatial Laplacian with eigenvalues between zero and $\frac{1}{6}$ times the spatial curvature. This last type of eigenfunctions cannot be square integrated, and they are responsible for long-range correlations in a spatially open universe [6]. In spite of the fact that these perturbations cannot be square integrated, they may naturally occur in an open universe which is created in an exponentially expanding false vacuum [7, 8], or they may be generated during preheating [9].

Another important property of the open Friedmann-Lemaître-Robertson-Walker (FLRW) geometry is that a spatial volume and the surface of its boundary grow at the *same* rate when the infinite volume limit is taken. The combination of large boundary surfaces and the presence of long-range correlations in open spacetimes appears to have an effect on the growth of surface terms at spatial infinity in these spacetimes.

Besides the theoretical reasons which make the open FLRW spacetime an interesting object to study, the open FLRW geometry has gained relevance as a model for the observed universe, with observations favoring a relatively small value of the density parameter [10]. Furthermore, progress has been made in describing the creation of an open FLRW universe from an exponentially expanding false vacuum (see, *e.g.*, [11, 12] and section 3 in the next chapter), and the theory of perturbations in open FLRW spacetimes has been worked out in greater detail [6] - [16].

This chapter is structured as follows. In section 3.2 we discuss the physical relevance of action-extremizing field configurations, and we show that surface terms can contribute to the variation of the action for square-integrable perturbations. In section 3.4 we decompose the scalar field perturbations in terms of eigenfunctions of the spatial Laplacian, and we discuss the occurrence of supercurvature modes. The dynamics of the extremal action configurations is considered in section 3.5, and we recover the usual equation of motion for each perturbation component, with an additional source term, which can be expressed in terms of a surface integral which is evaluated at spatial infinity. We show that this source term can be neglected in the case where there are only subcurvature excitations of the scalar field, but it appears to diverge in the case where there are supercurvature perturbations. Due

to this divergence, extremality of the action can only be defined in the restricted phase-space of field perturbations for which surface terms are finite. Depending on how one restricts the phase-space of field perturbations, a nontrivial source term contributes to the equation of motion for the extremal action configurations. In section 3.6, we consider the quantum correlation function of the scalar field. In the case where there are supercurvature perturbations, it is shown that the action functional is sensitive to degrees of freedom of the scalar field which have zero L_2-norm. It therefore appears that the correlation function is not well defined, unless one adopts nontrivial constraints on the phase-space of the scalar field, or one needs to include the zero-norm degrees of freedom in the integration over paths.

3.2 The extremal action principle

In this section we introduce the variational approach to classical field theory. We then use arguments from quantum field theory to motivate a modified form of the variational method in a cosmological context. Surprisingly, it appears that non-local interactions at a classical level can emerge from the underlying quantum theory with a standard expression for the Lagrangian. While our explicit calculations involve only the simple case of a scalar field, our arguments are relevant in a more general field theoretical context, including general relativity. We will come back to this point at the very end of this section.

One way of describing the dynamics of a classical field is by formulating a field equation. A specific solution of the field equation is determined by the boundary or periodicity conditions which apply to the system. It is of interest to note that the dynamics of the fields which can be observed in nature are described by field equations which act locally, while mathematical consistency does not require this. Hence, the dynamics of classical fields has a local aspect, in the sense that the field equations involve only the field variables and derivatives thereof at each point. Further, a particular classical field configuration is subject to global constraints, which act in the form of boundary or periodicity conditions. The work on this chapter started as an attempt to establish whether the local aspects of the dynamics of fields, which is apparent from the structure of the field equations, are fundamental in nature.

In order to gain a deeper insight in the global aspects of the dynamics of fields, a Lagrangian approach appears to be most suitable. In this approach, an action functional is constructed from the field variables over the entire spacetime. The dynamics of the field then follows by requiring that the action is extremal in the space of field configurations. Establishing extremality of the action amounts to showing that the action does not vary at first order, for arbitrary infinitesimal perturbations of the field variables. It is essential to note that the field perturbations which are used to 'test' the extremality of the action in the classical description, are purely a mathematical construct. Further, we stress that the choice of the action functional is motivated with the aim to recover the field equations, and hence the

Lagrangian description has the same physical content as the field equation.

At this point, let us formulate more precisely the question whether the local form of the interactions in nature is fundamental. On the one hand, it is well known that there exist conserved quantities which are related to global symmetries of the action [17] [18]. Although the existence of conserved quantities suggests an underlying global aspect of the dynamics of classical fields, this global aspect is in fact a consequence of applying Gauss's theorem to a four-divergence, which vanishes locally as at each point in our spacetime as a consequence of the field equation. On the other hand, there is the question whether there can be a non-local coupling between physical fields, which acts at the level of the field equation. In particular, one would like to know whether non-local interactions at a classical level can emerge from an underlying quantum theory for which the Lagrangian has the usual local form. In this chapter, we will focus on this last question.

Let us now consider in some detail how classical field theory arises as a limit of an underlying quantum field theory. According to the Feynman path-integral approach to quantum field theory, the expectation value of an operator O which acts on a field ψ, is given by the formal expression,

$$\langle O \rangle = Z^{-1} \int \mathrm{d}[\psi] O[\psi] \, e^{iS[\psi]/\hbar}, \tag{32}$$

where $S[\psi]$ is the action functional, Z is a normalization constant, and $\mathrm{d}[\psi]$ is a measure on the space of field configurations (see, *e.g.*, [19]). The integral is evaluated over all field configurations (paths) which are continuous and which satisfy certain initial or periodicity conditions. One should note that there is considerable difficulty involved in making the path-integral well defined, which is due to the fact that typical paths which contribute to the integral are non-differentiable. In our derivation, where we consider a free field, the different degrees of freedom decouple, and one can ignore those degrees of freedom which vary with infinite frequency.

As $\hbar$ approaches zero in expression (32), the oscillatory behavior of the integrand suggests that the integral is dominated by those field configurations ψ which are in some sense near to a field configuration ψ_0 which extremizes the action. Since $\hbar$ is close to zero when expressed in terms of macroscopic units of time and energy, one therefore expects that classical physics is accurately described by an action extremizing field configuration ψ_0. The essential difference between this classical limit, and the classical theory which we discussed previously, is the fact that in the former case there are physical field perturbations which probe the phase-space nearby an action extremizing configuration, while in the latter case these field perturbations are purely a mathematical construct. As we will show in the following, this difference can give rise to an essentially different expression which describes the dynamics of the classical field.

As is well known, extremality of the action for ψ_0 implies that this configuration satisfies the classical field equations, *provided* that a surface term vanishes for all

paths. In a classical variational treatment, surface terms are set to zero trivially by considering only paths which have compact support. However, this restriction on the type of paths does not occur in the sum over paths (32), and it seems natural to consider field configurations ψ_0 which extremize the action for the most general class of paths for which extremality of the action can be defined.

Indeed, one should note that in a classical treatment of cosmological perturbations one does not normally assume that perturbations must have compact support. However, if one accepts that classical perturbations do not have compact support, then it seems rather unnatural to require that quantum fluctuations about the classical field configurations have compact support. If this would be the case, then there would be a finite distance beyond which there are still classical perturbations while quantum fluctuations vanish. This appears to contradict the Copernican principle, which is commonly adopted in cosmology.

Considering the relation between classical and quantum physics, it should be mentioned that the path-integral approach does not only explain more than a classical approach (*i.e.*, testable quantum effects), but one also needs to assume more than in classical physics (*e.g.*, the existence of a classical regime [20], as well as various infinite subtractions [19]). One might therefore feel that the validity of the path-integral approach is as questionable as the classical variational approach, when it is applied to cosmological situations where it has not been tested. When seen in this light, the classical assumption that field-perturbations are restricted to have compact support is not proven to be wrong, but rather, it represents one possible choice in a more general class of boundary or asymptotic conditions. Whichever point of view one favors, it seems interesting to investigate the implications of relaxing the assumption that field-perturbations must have compact support. We will discuss these implications in the following.

3.3 Scalar field in FLRW geometry

The line element of the FLRW geometry is given by,

$$\mathrm{d}s^2 = -\mathrm{d}t^2 \ + a^2(t) \left[\mathrm{d}\chi^2 \ + c^{-2} \sinh^2 c\chi (\mathrm{d}\theta^2 + \sin^2\theta \mathrm{d}\phi^2) \right], \qquad (33)$$

where $c = \mathbf{R}^+$ for the spatially open geometry, while the spatially flat and closed geometry are obtained by taking the limit $c \downarrow 0$ or by choosing $c \in i \times \mathbf{R}^+$ respectively. We will refer to the geometry with the line element (33) as $\mathcal{M}$, while a spatial hypersurface of constant time t is referred to as Σ.

It follows directly from expression (33) that the surface of a spatial sphere of constant radius χ_0 grows as fast as the three-volume inside the sphere, when one considers the limit where $\chi_0 \to \infty$. One may therefore expect that surface terms can be equally important as volume terms when we take the infinite volume limit in an open universe. This situation is essentially different from the situation in a spatially flat spacetime, where the surface of a spatial sphere of constant radius χ

grows by one power of χ less fast then the three-volume which is contained inside the sphere.

We will consider a scalar field ψ, which is described by the Lagrangian density

$$\mathcal{L}[\psi] = -\frac{1}{2}\sqrt{-g}\ (g^{\mu\nu}\partial_\mu\psi\partial_\nu\psi\ + m^2\psi\psi), \tag{34}$$

where $g_{\mu\nu}$ denotes the FLRW metric (33), and $g = \det(g_{\mu\nu})$.

We define the action of the ψ-field as the integral of the Lagrangian density (34) over the entire spacetime,

$$S[\psi] := \int \mathrm{d}^4x\mathcal{L}[\psi]. \tag{35}$$

Note that the integral in this expression does not need to converge. This is not necessarily a problem if one is interested in calculating the variation of the action under a change of the field from ψ to $\psi + \delta\psi$, where $\delta\psi$ is a suitably small 'test-perturbation'. The question arises which restriction one has to impose on the test-perturbations $\delta\psi$ so that the first-order variation δS is well defined. The first-order variation of the action (35) follows by the standard procedure of functional derivation,

$$\delta S = \int \mathrm{d}^4x \left(\frac{\delta\mathcal{L}}{\delta\psi}\delta\psi + \frac{\delta\mathcal{L}}{\delta\partial_\mu\psi}\delta\partial_\mu\psi\right). \tag{36}$$

By partially integrating equation (36), where the Lagrangian is given by expression (34), we obtain

$$\delta S = \int \mathrm{d}^4x\sqrt{-g}\left[\psi^{;\mu}_{;\mu} - m^2\psi\right]\delta\psi\ - \int \mathrm{d}^4x\sqrt{-g}(\psi^{;\mu}\delta\psi)_{;\mu}\ , \tag{37}$$

where a semicolon denotes the covariant derivative.

Provided that the second term on the right-hand side of equation (37) vanishes for nonzero perturbations $\delta\psi$, then the condition $\delta S = 0$ implies the vanishing of the term in brackets, and hence the field equation holds. This is the case when we consider test-perturbations $\delta\psi \in D$, where D is defined as the class of perturbations which are bounded and which have compact support. However, as we mentioned in the beginning of this section, the restriction to test-perturbations $\delta\psi \in D$ does not follow from known physical principles, when the spacetime itself is non-compact. Let us therefore try to determine the largest class of test-perturbations for which the variation of the action is well defined. For a scalar field ψ, and a Lagrangian which is bi-linear in the field variable, it is clear that square integrability of $\delta\psi$ is a necessary condition for the existence of the variation of the action (37), *i.e.*, we require $\delta\psi \in L_2(\mathcal{M})$. It is not *a priori* clear whether $\delta\psi \in L_2(\mathcal{M})$ is a sufficient condition for the existence of the variation of the action (37), and it may be necessary to restrict the type of test-perturbations further to ensure that δS exists. Assuming that we are able to determine the largest class of test-perturbations $\delta\psi$

for which δS exists, then it remains a question whether there exist field configurations ψ_0 such that δS vanishes for all perturbations $\delta\psi$ about ψ_0.

Let us first address the question whether the restriction $\delta\psi \in L_2(\mathcal{M})$ is sufficient to ensure the existence of δS. The answer to this question is negative, which we show by an example where the contribution of surface terms to δS diverges, while ψ is a solution of the field equation and $\delta\psi \in L_2(\mathcal{M})$. Since we will focus on surface effects at *spatial* infinity, we require that $\delta\psi$ can be square integrated over a spatial hypersurface of constant time in the geometry (33), *i.e.*, $\delta\psi \in L_2(\Sigma)$, while we do not specify the time dependence of $\delta\psi$. It is clear from the expression of the line element (33) that a square integrable test-perturbation $\delta\psi$ must approach zero faster than $1/\chi$ in the spatially flat case, and faster than $e^{-\chi}$ in the spatially open case. A specific example of a square integrable test-perturbation is given by

$$\delta\psi = (1+\chi)^{-(1+\alpha)}\partial_\chi\psi \qquad \text{and} \qquad \delta\psi = e^{-(1+\alpha)\chi}\partial_\chi\psi, \tag{38}$$

in the spatially flat and open case respectively, and $\alpha \in \mathbf{R}^+$. By substituting expressions (38) for $\delta\psi$ into equation (37), and using (33), we find

$$\delta S = -4\pi a^{-2}(t) \int \mathrm{d}t\mathrm{d}\Omega \lim_{\chi\to\infty} F(\chi)\,(\partial_\chi\psi)^2, \tag{39}$$

where $\mathrm{d}\Omega$ denotes the volume element on the unit two-sphere, and $F(\chi) = \chi^{1-\alpha}$ in the spatially flat case, and $F(\chi) = e^{(1-\alpha)\chi}$ in the spatially open case. Indeed, expression (39) diverges for some values of $\alpha \in (0,1]$, provided that the term $\partial_\chi\psi$ does not approach to zero as fast as $F^{-1/2}(\chi)$ in the limit where $\chi \to \infty$. The variation of the action (39) can therefore be arbitrarily large, for $\delta\psi \in L_2(\Sigma)$.

Let us now address the question whether there exist configurations of the ψ-field which extremize the action for all $\delta\psi \in L_2(\Sigma)$, in the cosmologically interesting case where ψ and $\partial_\chi\psi$ do not vanish at spatial infinity. We show that the answer to this question is negative. We will therefore use a result which is derived in the following, which states that a field configuration which extremizes the action for all $\delta\psi \in L_2(\Sigma)$ must be a solution of the field equation. We combine this with the result which was derived earlier in this section, which shows that a solution of the field equation for which $\partial_\chi\psi$ does not approach to zero at spatial infinity, does not extremize the action for all $\delta\psi \in L_2(\Sigma)$. Hence, it follows that action extremizing configurations do not exist for $\delta\psi \in L_2(\Sigma)$ and $\partial_\chi\psi$ not approaching to zero at infinity.

In deriving the proof above, we assumed that a field configuration which extremizes the action for all $\delta\psi \in L_2(\Sigma)$ must be a solution of the field equation. In order to proof this, let us recall that for $\delta\psi \in D$, *i.e.*, the class of test-perturbations which are bounded and which have compact support, extremality of the action implies that the field equation holds and vice-versa. Configurations which do not satisfy the field equation can therefore not extremize the action for all $\delta\psi \in D$, and since $D \subset L_2(\Sigma)$ these configurations do not extremize the action for all

$\delta\psi \in L_2(\Sigma)$. Hence it follows that a field configuration which extremizes the action for all $\delta\psi \in L_2(\Sigma)$ must be a solution of the field equation, which proves our assumption.

The observation that action extremizing configurations do not in general exist for $\delta\psi \in L_2(\Sigma)$, implies that the usual identification between classical physics and action extremizing configurations becomes ambiguous when we allow for perturbations which do not fall off sufficiently fast at infinity. There are several ways by which one could try to resolve the problem which is posed by the non-existence of extremal action configurations for test-perturbations $\delta\psi \in L_2(\Sigma)$. We will discuss these possible solutions in the following.

First, let us recall that the restriction $\delta\psi \in L_2(\Sigma)$ was found to be *necessary* to ensure finiteness of δS, but due to the contribution of a surface term to δS this restriction is not *sufficient*. This observation suggests that the class of test-perturbations $\delta\psi$ should be restricted further, such that δS is finite for all $\delta\psi$. Although finiteness of δS is easily achieved by requiring that the test-perturbations $\delta\psi$ fall off sufficiently fast, this does not imply that extremal action configurations exist in the space of test-perturbations for which δS is finite. The reason for this is that the existence of extremal action configurations requires that the surface term contribution to δS vanishes completely, which is clearly a stronger restriction on $\delta\psi$ than the condition that δS is finite. Although one could restrict $\delta\psi$ to ensure that the surface term contribution to δS vanishes completely, this would be rather add-hoc since it is not shown that this is the only possible restriction on the class of test-perturbations for which extremal action configurations exist.

Instead of restricting the class of test-perturbations, one could also attempt to remove the contribution of surface terms to δS by modifying the Lagrangian density (34). Let us therefore note that the choice of the Lagrangian density is motivated by the fact that one recovers the Klein-Gordon equation, provided that the variation of the action and the surface term in equation (37) vanish. In the classical variational approach, where surface terms are made to vanish by assuming boundary conditions on $\delta\psi$, one therefore has the freedom to add a term to the Lagrangian density which has the form of a four-divergence, since the variation of this term equals a vanishing surface term. In this section we questioned the assumption that the surface term in equation (37) vanishes in perturbed flat and open FLRW spacetimes. However, it is conceivable that one can add a four-divergence term to the Lagrangian (34) such that its variation cancels the surface term in equation (37). Indeed, in the context of Hamiltonian cosmology, as well as in quantum cosmology, it appears to be natural to add a surface term to the Einstein-Hilbert action which has the property that its variation cancels an identical term which arises from the variation of the Einstein-Hilbert action [21] - [23].

Let us now consider whether the same possibility exists in the case where we are dealing with a scalar field. We therefore add a generic surface term to the

action (35), which has the form

$$S_B[\psi] = \frac{1}{2}\int \mathrm{d}^4x\sqrt{-g}B^{\mu}_{;\mu}\,, \tag{40}$$

where $B^\mu = B^\mu[\psi]$, and then we consider whether the variation of this surface term may cancel the surface term in equation (37). The variation of B^μ follows by the method of functional derivation, *i.e.*, treating ψ and $\partial_\nu\psi$ as independent variables:

$$\delta B^\mu = \frac{\delta B^\mu}{\delta\psi}\delta\psi \; + \frac{\delta B^\mu}{\delta\partial_\nu\psi}\delta\partial_\nu\psi, \tag{41}$$

where we used that B^μ cannot depend on higher than first-order derivatives of ψ. It is clear that any dependence of B^μ on higher than first-order derivatives of ψ contributes terms to the variation of the action which are proportional to the variation of higher than first-order derivatives of $\delta\psi$. These terms cannot cancel against the surface term in equation (37), which contains at most first-order derivatives of $\delta\psi$, although a cancellation was required.

The requirement that the surface term in equation (37) cancels the surface term which arises from the variation of S_B results in the conditions

$$\frac{\delta B^\mu}{\delta\psi} = \psi^{;\mu}, \quad \text{and} \quad \frac{\delta B^\mu}{\delta\partial_\nu\psi} = 0, \tag{42}$$

for all μ,ν, and we used expression (41). The first condition in equation (42) constrains B^μ to be of the form $B^\mu = \psi^{;\mu}\psi + \mathrm{c}_1$, where c_1 is a functional which does not depend on ψ, while the second condition constrains B^μ to be a functional which does not depend on $\partial_\mu\psi$. Clearly, both requirements are exclusive, and there exists no functional B^μ such that the variation of S_B, (40), cancels the surface term in equation (37). Note, however, that the precise form of the surface term in equation (37) does change by adding a term of the form (40) to the action. Hence, the contribution of a surface term to the variation of the scalar field action (35) appears to be generic, although its precise form is ambiguous. In the following calculation we will retain the surface term which appears in equation (37), which means that we assume S_B to vanish.

Having considered the possibility to adopt further restrictions on the type of test-perturbations, as well as modifying the action by adding a surface term contribution, we have not found an argument which shows us that we can neglect the contribution of a surface term to the variation of the action. However, taking the surface term in equation (37) seriously confronts us with the problem that field configurations which extremize the action in the space of test-perturbations for which δS is well defined, do not in general exist. It should be noted, however, that the non-existence of action extremizing field configurations does not need to be a problem if one could show that those test-perturbations for which the action functional is not extremal, have a zero phase-space measure in the space of fields

ψ. Indeed, it is clear that paths of the form (39), which yield large surface terms at spatial infinity, are highly special in the sense that the asymptotic behavior of these paths is correlated with the field ψ about which we expand. Therefore, one expects that these paths occupy a very small amount of phase-space in the space of field configurations in which extremality of the action is considered, and their relevance for the dynamics of the ψ-field may be negligible. Note, however, that precisely the same argument applies to the case where $\delta\psi \in D$, since in this case $\delta\psi$ is specified to be exactly equal to zero for arbitrarily large radii χ. In order to make these considerations quantitative, it is necessary to introduce a measure on the phase-space of the ψ-field. We will address this problem in the following sections.

3.4 Perturbations in open FLRW

In order to obtain a quantitative description of the space of field configurations of the scalar field ψ, it is useful to decompose ψ and test-perturbations $\delta\psi$ in terms of eigenfunctions of the spatial Laplacian which are complete in the space L_2 of functions which are square integrable on the hypersurfaces $\Sigma(t)$. The reason why it is convenient to use eigenfunctions of the spatial Laplacian, is that this operator is present in the expression for the variation of the action (37). When we ignore the surface term, it is therefore clear that each eigenfunction only couples to itself, and the dynamics of each mode is independent of the dynamics of all other modes.

Let $Q(x)$ be a solution of the Helmholtz equation, *i.e.*,

$$Q^{;i}_{;i} + (k/a)^2 Q = 0, \tag{43}$$

where $;i$ denotes the covariant derivative with respect to the coordinate $x^i \in \{r, \theta, \phi\}$ in the geometry (33), $a = a(t)$ denotes the scale factor, and $k \in \mathbf{R}^+$. In the following, we concentrate on the spatially open geometry (33), while we consider the spatially flat spacetime as a limiting case of the spatially open geometry. A basis of solutions of equation (43), which are complete in the space of L_2 functions on $\Sigma(t)$, and which factorize in terms of an angular and a radially dependent part, is given by

$$Z_{qlm} = \Pi_{ql}(\chi)\ Y_{lm}(\theta, \phi), \tag{44}$$

where Y_{lm} are the standard spherical harmonics on the unit two-sphere, and the radially dependent functions $\Pi_{ql}(\chi)$ are solutions of the equation

$$\frac{1}{g_2}\frac{\partial}{\partial\chi} g_2 \frac{\partial}{\partial\chi}\Pi_{ql}(\chi) = \left(k^2 - \frac{l(l+1)}{g_2}\right)\Pi_{ql}(\chi), \tag{45}$$

where $g_2 = c^{-2}\sinh^2 c\chi$. Equation (45) has solutions of the form,

$$\Pi_{ql}(\chi) = N_{ql}(\sinh c\chi)^l \left(\frac{-1}{\sinh c\chi}\frac{\mathrm{d}}{\mathrm{d}\chi}\right)^{l+1} \cos(qc\chi), \tag{46}$$

where q is defined by $q^2 = k^2/c^2 - 1$, and

$$N_{ql} := \sqrt{\frac{2}{\pi}} \left[\prod_{n=0}^{l} (n^2 + q^2) \right]^{-1/2} \tag{47}$$

is a normalization factor [24, 25]. Notice that the $q = 0$ mode solves the Helmholtz equation (43) with a *nonzero* eigenvalue equal to $-c^2/a^2$, which equals $\frac{1}{6}$ times the spatial curvature in the geometry (33).

The radial solutions for the spatially flat geometry are obtained by taking the limit $c \downarrow 0$ in expression (46), keeping k fixed,

$$\lim_{c \downarrow 0} \Pi_{ql}(\chi) = \sqrt{\frac{2}{\pi}} k j_l(k\chi), \tag{48}$$

where j_l denotes the spherical Bessel function [30]. From now on, we assume that the spacetime is open, such that $c \in \mathbf{R}^+$, and without loss of generality we may set $c = 1$ in expression (33) by absorbing a factor c in the definition of the comoving radial coordinate χ and by absorbing a factor c^{-1} in the definition of the scale factor $a(t)$.

It follows from expression (46) that the radial functions Π_{ql} can be written as the product of an oscillating factor $\cos q\chi$ or $\sin q\chi$, and a factor which approaches to zero exponentially as $\sinh^{-1}\chi$ in the limit where $\chi \to \infty$. Since the modes Z_{qlm} with $q \in \mathbf{R}^+$ vary at comoving length scales which are typically smaller than the curvature scale which we have set equal to one in the FLRW geometry (33), these modes are called *subcurvature modes.*

There exist solutions of the Helmholtz equation (43) for which $k^2 \in (0, 1]$, which corresponds to imaginary values of $q \in i \times (0, 1]$. The explicit expression for these modes is still given by equation (46), where the factor $\cos(q\chi)$ is replaced by $\cosh(|q|\chi)$. The modes Z_{qlm} with $q \in i \times (0, 1]$ approach to zero as a constant times $\exp((|q| - 1)\chi)$ in the limit where $\chi \to \infty$, and since they vary at length scales greater than the curvature scale one calls them *supercurvature modes.*

We define the spatial integration operation by

$$\langle f \rangle := \lim_{\epsilon \downarrow 0} \langle f \rangle(\epsilon) \tag{49}$$

where

$$\langle f \rangle(\epsilon) := \int \mathrm{d}\Omega \int_0^{1/\epsilon} \mathrm{d}\chi \sinh^2(\chi) f. \tag{50}$$

and $\mathrm{d}\Omega^2$ denotes the volume element on the unit two-sphere. The subcurvature modes $Z_{qlm}(q \in \mathbf{R}^+)$ are orthonormal with respect to spatial integration,

$$\langle Z_{qlm} Z_{q'l'm'} \rangle = \delta(q - q')\delta_{ll'}\delta_{mm'}, \tag{51}$$

and they are known to be complete in the space $L_2(\Sigma)$ [5], which consists of equivalence classes of functions f for which $\langle |f|^2 \rangle$ exists, where we identify functions f which differ only on a set of Lebesgue measure zero.

For the supercurvature modes, the indefinite integral over the radius in expression (51) does not exist, so that these modes cannot be normalized in the $L_2(\Sigma)$ sense. Furthermore, expression (51) diverges when only one of the modes Z corresponds to a supercurvature mode, and $l = l'$ and $m = m'$. Therefore, the supercurvature modes cannot be decomposed in terms of the subcurvature modes. Mechanisms which may be responsible for the generation of supercurvature perturbations in open spacetimes have been investigated in [14, 8].

The ψ-field may be expanded in terms of the modes Z_{qlm},

$$\psi(x,t) = \psi^-(x,t) \ + \psi^+(x,t), \tag{52}$$

where

$$\psi^-(x,t) := \sum_{lm} \int_0^\infty \mathrm{d}q \ \psi_{qlm}(t) Z_{qlm}(x), \tag{53}$$

$$\psi^+(x,t) := \sum_{lm} \int_0^i \mathrm{d}\bar{q} \ \psi_{\bar{q}lm}(t) Z_{\bar{q}lm}(x), \tag{54}$$

where $x = \{\chi, \theta, \phi\}$, and the integration over $\bar{q}$ runs along the imaginary axis in the complex $\bar{q}$-plane.

An important class of perturbations, which is believed to occur in the early universe, corresponds to the case where the coefficient of each independent mode is chosen according to a Gaussian probability distribution (see, *e.g.*, [26] - [28]). For this type of perturbation, which is called a 'Gaussian perturbation' or 'random-field', there are no correlations between the coefficients ψ_{qlm} for different values of q, l, and m. The statistical properties of a random-field are determined by the variance of the Gaussian probability distribution, which we call σ. In the generic case, where σ depends on q, l, and m, one cannot determine the variances $\sigma(q,l,m)$ from a single realization of a random-field, which is determined by the set of coefficients ψ_{qlm}. Instead, one would need an infinite *ensemble* of random-fields, in order to deduce the statistical properties, *i.e.*, the variances $\sigma(q,l,m)$, according to which these random-fields are generated.

Let us now define the *ensemble average* of a functional as the weighted sum of this functional over all random-fields in an ensemble, where the weight factor is given by the probability for each specific random-field to occur. This allows us to define the two-point correlation function of the ψ-field as the ensemble average of $\psi(x)$ times $\psi(x')$. A random-field $\psi(x)$ is said to be statistically homogeneous and isotropic when the two-point correlation function is invariant under the group of isometries on Σ, *i.e.*, the group of rotations and spatial translations. Clearly, the two-point correlation function of a statistically homogeneous and isotropic random field can only be a function of a distance measure which is invariant under the group of isometries on Σ, and we can take this distance measure to be the length $d(x,x')$ of a geodesic which relates the points x and x'. In can be shown that statistical homogeneity and isotropy of a random-field $\psi(x)$ holds if and only if the variances σ do not depend on the labels l and m [29].

Although it seems rather artificial to introduce the concept of an ensemble in the context of cosmology, since we can only observe one universe, a physical interpretation of the ensemble average is provided by the property of *ergodicity*. In the context of random-fields, ergodicity is defined as the equivalence of ensemble averaging and spatial averaging, where the spatial average of the two-point correlation function is defined by summing $\psi(x)$ times $\psi(x')$ over random sets of points x and x' for which the geodesic distance $d(x,x')$ has a specific value. In the case where Σ is a Euclidean three-space, ergodicity can be proven to hold under fairly weak assumptions [26], but for a hyperbolic three-space no proof seems to be known, while it is usually assumed.

In the following, we will assume a Gaussian statistically homogeneous and isotropic spectrum of subcurvature perturbations. One should note that this type of perturbation cannot be square integrated. This follows by substituting the expansion of ψ, (53), into the hypersurface integral (49) and using the orthonormality relation (51). The resulting expression contains an indefinite sum over l and m of the squared coefficient ψ_{qlm}, and this sum diverges when the variance $\sigma(q)$ is nonzero. It is therefore clear that the property of non-square integrability is not specifically related to the presence of supercurvature modes.

3.5 Extremal action dynamics

Let us now calculate the variation of the action (37), which is evaluated over a bounded spatial volume $V(\chi_0)$, which we define as those points in the geometry (33) for which $\chi < \chi_0$, and then we consider the limit where $\chi_0 \to \infty$. We obtain

$$\delta S = \int \mathrm{d}t \lim_{\chi_0 \to \infty} \left[a^3 \int \mathrm{d}\Omega^2 \int_0^{\chi_0} \mathrm{d}\chi \, \sinh^2\chi \, \delta\psi \left(\frac{1}{\sqrt{-g}} \partial_\mu g^{\mu\nu} \sqrt{-g} \, \partial_\nu \, - m^2 \right) \psi \right.$$

$$\left. - a \sinh^2\chi \int \mathrm{d}\Omega^2 \, \delta\psi \partial_\chi \psi \Big|_{\chi=\chi_0} \right], \tag{55}$$

where $a = a(t)$. Using the definition of the integration operation (51), expression (55) can be written in the form,

$$\delta S = \int \mathrm{d}t \left[a^3 \left\langle \delta\psi \, \left(\frac{1}{\sqrt{-g}} \partial_\mu g^{\mu\nu} \sqrt{-g} \partial_\nu - m^2 \right) \, \psi \right\rangle \right.$$

$$\left. - \, a \lim_{\chi_0 \to \infty} \sinh^2\chi_0 \int \mathrm{d}\Omega \delta\psi \partial_\chi \psi \, \Big|_{\chi=\chi_0} \right]. \tag{56}$$

We will consider separately the cases where the expansion of the field ψ includes only subcurvature modes, and the case where the expansion includes supercurvature modes as well.

3.5.1 Open spacetime with subcurvature perturbations

Let us first consider the case where the field ψ can be expanded in terms of only subcurvature modes, *i.e.*, we assume that $\psi_{qlm} = 0$ for all $q \in i \times (0, 1]$, so that only the first term in the expansion of the field (52) is nonzero. Equation (56) can then be evaluated separately for each mode, by substituting the expansion (52) into expression (56), and using the orthonormality relation (51). We obtain

$$\delta S = \int \mathrm{d}t \int \mathrm{d}q \sum_{l,m} \delta\psi_{qlm}(t) \left[a^3 \left(\frac{1}{\sqrt{-g}} \partial_0 g^{00} \sqrt{-g} \partial_0 - a^{-2}(t) k^2 - m^2 \right) \psi_{qlm}(t) \right. \tag{57}$$

$$\left. - \lim_{\chi_0 \to \infty} a \sinh^2 \chi_0 \int \mathrm{d}q' \psi_{q'lm} \, \Pi_{ql} \partial_\chi \Pi_{q'l} \Big|_{\chi=\chi_0} \right].$$

The requirement that the variation of the action vanishes for nonzero perturbations $\delta\psi_{qlm}(t)$ implies an equation of motion for each perturbation component $\psi_{qlm}(t)$, namely,

$$\left(\frac{1}{\sqrt{-g}} \partial_0 g^{00} \sqrt{-g} \partial_0 - a^{-2} k^2 - m^2 \right) \psi_{qlm}(t) = J_{qlm}, \tag{58}$$

where

$$J_{qlm} := \lim_{\chi_0 \to \infty} \left[a^{-2} \sinh^2 \chi_0 \int \mathrm{d}q' \psi_{q'lm} \, \Pi_{ql} \partial_\chi \Pi_{q'l} \Big|_{\chi=\chi_0} \right]. \tag{59}$$

Note that J_{qlm} acts as a *source term* in equation (58), and this term couples perturbations which have the same angular wave numbers l and m. One would like to know whether the limit in expression (59) exists, and whether or not this term can be neglected. In order to answer this question, we need to evaluate the integral over q' of the distribution $\psi_{q'lm}$, which is multiplied by a factor which is of order unity. According to equation (51) and (52), the distribution $\psi_{q'lm}$ can be defined by,

$$\psi_{q'lm} = \lim_{\epsilon \downarrow 0} \psi_{q'lm}(\epsilon), \tag{60}$$

where

$$\psi_{q'lm}(\epsilon) := \langle Z_{q'lm} \psi \rangle(\epsilon) \tag{61}$$

and the limit $\epsilon \downarrow 0$ should be evaluated after the integration over q' is performed. When we integrate over a bounded volume, then the modes Z_{qlm} are dependent in the sense that their overlap $\langle Z_{qlm} Z_{q'lm} \rangle(\epsilon)$ is nonzero and of the order of ϵ^{-1} for $q - q'$ of the order of ϵ. The number of independent modes in a fixed q'-interval therefore tends to diverge as ϵ^{-1} in the limit where $\epsilon \downarrow 0$. In the previous section, we introduced the concept of a Gaussian perturbation. In order to generate a Gaussian perturbation which has an amplitude of order one, the coefficients ψ_{qlm} in the expansion of the field (53) need to be uncorrelated for values of q differing more than ϵ, while the amplitude of the coefficients must diverge as $\epsilon^{-\frac{1}{2}}$ when $\epsilon \downarrow 0$. The asymptotic behavior of the integral over q' in expression (59) can therefore be

estimated as the sum of ϵ^{-1} uncorrelated numbers which are of the order of $\epsilon^{-1/2}$, multiplied by a q'-interval which is of the order of ϵ. In the limit where $\epsilon \downarrow 0$, the term between brackets in expression (59) will therefore remain of order one, and the expression does not converge. Note, however, that the left-hand side of the equation of motion (58) is proportional to the coefficient ψ_{qlm}, which diverges as $\epsilon^{-1/2}$ in the limit where $\epsilon \downarrow 0$. We therefore find that the source term on the right-hand side of equation (58) can be neglected in the infinite volume limit, when the perturbations of the field are Gaussian and of the subcurvature type.

3.5.2 Open spacetime with supercurvature perturbations

Let us now attempt to derive an equation of motion for the ψ-field, in the case where the expansion of the ψ-field (52) includes supercurvature perturbations.

We may therefore substitute the expansion of the ψ-field (52) in the expression for the variation of the action (55), which yields,

$$\delta S = \int \mathrm{d}t \lim_{\chi_0 \to \infty} \tag{62}$$

$$\times \left[a^3 \int \mathrm{d}\Omega^2 \int_0^{\chi_0} \mathrm{d}\chi \, \sinh^2\chi \, \delta\psi \left(\frac{1}{\sqrt{-g}} \, \partial_\mu g^{\mu\nu} \sqrt{-g} \, \partial_\nu \, - m^2 \right) (\psi^- + \psi^+) \right.$$

$$\left. -a \int \mathrm{d}\Omega^2 \sinh^2\chi \, \delta\psi \partial_\chi (\psi^- + \psi^+) \, \Big|_{\chi=\chi_0} \right].$$

Using the definition of the integration operation (49), and expression (52), we recover expression (58), with an additional source term which accounts for the coupling between subcurvature and supercurvature perturbations, *i.e.*,

$$\left(\frac{1}{\sqrt{-g}} \partial_0 g^{00} \sqrt{-g} \partial_0 \, - a^{-2}k^2 - m^2 \right) \psi^-_{qlm}(t) = \, J_{qlm} + J^+_{qlm}, \tag{63}$$

where $q \in \mathbf{R}^+$, J_{qlm} is given by expression (59), and

$$J^+_{qlm} := \lim_{\chi_0 \to \infty} \int_0^i \mathrm{d}\bar{q} \left[\left(\frac{1}{\sqrt{-g}} \partial_0 g^{00} \sqrt{-g} \partial_0 \, - a^{-2}k^2 - m^2 \right) \psi^+_{\bar{q}lm}(t) \right.$$

$$\left. \times \int_0^{\chi_0} \mathrm{d}\chi \sinh^2\chi \Pi_{ql} \Pi_{\bar{q}l} \, + a^{-2} \sinh^2\chi_0 \, \psi^+_{\bar{q}lm} \, \Pi_{ql} \partial_\chi \Pi_{\bar{q}l} \Big|_{\chi=\chi_0} \right]. \tag{64}$$

Note that both terms which contribute to expression (64) diverge exponentially in the limit where $\chi_0 \to \infty$, and the limit in this expression does not exist, unless the divergent terms cancel. Let us therefore observe that the two terms at the right-hand side of equation (64) diverge exponentially as $\exp|\bar{q}\chi|$, (see section 3.4), and both terms oscillate due to the radial function Π_{qlm}. A cancellation of the divergent

terms in equation (64) requires that both terms oscillate with the same phase. By re-writing equation (64), using,

$$\int_0^{\chi_0} \mathrm{d}\chi \sinh\chi \; \Pi_{ql}\Pi_{\bar{q}l} = \frac{\sinh^2\chi_0}{q^2-\bar{q}^2} \mid \Pi_{ql}\partial_\chi\Pi_{\bar{q}l} \; - \Pi_{\bar{q}l}\partial_\chi\Pi_{ql} \mid_{\chi=\chi_0}, \tag{65}$$

one finds that J^+_{qlm} diverges as the product of an exponential factor $\exp(|\bar{q}| + 1)\chi$, multiplied by the sum of two terms which oscillate out of phase as Π_{ql} and $\partial_\chi\Pi_{ql}$, respectively. Therefore, the right-hand side of equation (63) diverges, and we cannot use this equation to describe the time-evolution of the perturbation component $\psi_{qlm}(t)$. Recall that in the absence of supercurvature perturbations, surface terms appeared to give rise to a negligible correction to the equation of motion for each perturbation component $\psi_{qlm}(t)$, which followed by requiring that $\delta S = 0$ for all $\delta\psi \in L_2(\Sigma)$. When supercurvature perturbations are present, equations (62) and (63) show that it is precisely a surface term which contributes a divergent term to the variation of the action for all $\delta\psi \propto Z_{qlm}$. In this case, the extremal action condition $\delta S = 0$ cannot be satisfied for all $\delta\psi \in L_2(\Sigma)$, irrespectively of the equation of motion which the field satisfies. It is however clear that the condition $\delta S = 0$ must have solutions when test-perturbations are confined to some subspace of $L_2(\Sigma)$ for which δS is well defined. We will determine these subspaces in the following.

According to expressions (56) and (52), the surface term which contributes to δS behaves asymptotically as $\delta\psi$ times a factor $\sinh^2\chi\partial_\chi\psi^+$ in the limit where $\chi \to \infty$. The contribution of surface terms to the variation of the action (55) will therefore be finite and convergent, provided that $\sinh^2\chi\delta\psi\partial_\chi\psi^+$ converges when $\chi \to \infty$. Let us now *define* the class of test-perturbations $\{\delta\psi\}_c$ by the requirement that $\sinh^2\chi\delta\psi\partial_\chi\psi^+$ converges to a constant $c \in \mathbf{R}$ when $\chi \to \infty$.

Note that it follows from the definition of $\{\delta\psi\}_c$ that $\{\delta\psi\}_c$ contains D, i.e., the class of functions which are bounded and which have compact support. As is well known, the class of functions D is infinite dimensional in the sense that there exists a denumerable infinite set of linearly independent basis-functions which is complete in D [31], and therefore $\{\delta\psi\}_c$ must be infinite dimensional, for arbitrary $c \in \mathbf{R}$. It is therefore not clear whether one class of test-perturbations $\{\delta\psi\}_c$ for some specific value of $c \in \mathbf{R}$ dominates in terms of the phase-space which is occupied by these test-perturbations. We will make this statement more precise in the following section, where it is shown that that the classes of test-perturbations $\{\delta\psi\}_c$, for different values of $c \in \mathbf{R}$, are equivalent up to variations with vanishing $L_2(\mathcal{M})$-norm.

Summarizing, we found that the contribution of surface terms to the variation of the action diverges for square integrable field perturbations which do not fall off at a specific rate, depending on the spectrum of supercurvature perturbations. In the presence of supercurvature perturbations, extremality of the action can therefore only be defined with respect to a restricted class of field perturbations.

Surface terms contribute a non-trivial source term to the standard Klein-Gordon equation, but the magnitude thereof depends on the choice of the restricted class of test-perturbation with respect to which the action is extremized. The dynamics of the 'classical' field configurations therefore remains undetermined, unless one finds a physical argument which constrains the phase-space of the ψ-field uniquely.

3.6 Quantum correlations

In the previous section we showed that surface terms constrain the phase-space of test-perturbations for which the variation of the Klein-Gordon action is finite, in an open FLRW spacetime with supercurvature perturbations. One may also question whether the nontrivial surface terms which we found have an effect on quantum correlations of the ψ-field. As is clear from expression (32), the quantum correlation function of the ψ-field can be expressed as a weighted integral over all continuous field configurations, and the weight factor depends on the source term J^+, which may be infinite.

The two-point correlation function is given by the formal expression (see, *e.g.*, [19])

$$\tau(x, x') := Z^{-1} \int \mathrm{d}[\psi]\ \psi(x)\psi(x')\ e^{iS[\psi]/\hbar}, \tag{66}$$

where x denotes the set of coordinates on $\mathcal{M}$.

The standard method to calculate the two-point correlation function is to expand the field ψ, about some background configuration ψ_0, in terms of a denumerable complete set of solutions of the four-dimensional Helmholtz equation (see, *e.g.*, [32] for the details involved in this calculation). Since $L_2(\mathcal{M})$ is known to be separable, there exists a denumerable and complete set of solutions, which we call ψ_i, and we can choose these solutions to be orthonormal in $L_2(\mathcal{M})$. A generic expansion of the field ψ, about a configuration ψ_0, takes the form

$$\delta\psi := \psi - \psi_0 = \sum_i a_i \psi_i, \tag{67}$$

where $a_i \in \mathbf{R}$. Further, the measure on the space of the field ψ can be expressed in terms of the coefficients a_i, *i.e.*,

$$\mathrm{d}[\psi] = \prod_i \mu \mathrm{d}a_i, \tag{68}$$

where μ is a normalization constant with the dimension of inverse length, and the indefinite product runs over all values of the label i.

By substituting the expansions of the field (67) and the measure (68) into the expression for the correlation function (66), the path-integral can be evaluated explicitly. Assuming that there are no nontrivial source terms of the kind which we discussed in the previous section, then the standard expression for the two-point correlation function follows in terms of the complete set of modes ψ_i. We will not

repeat this calculation here, which can be found, *e.g.*, in [32], but instead we will consider what is the effect on the two-point correlation function (66) when there is a nontrivial source term $J^+[\psi]$ which contributes to the variation of the action.

Let us now define the set of functions $\tilde{\psi}_i \in \{\delta\psi\}_0$, which satisfy the property that the linear span of the modes $\tilde{\psi}_i$ is dense in $\{\delta\psi\}_0$, and the modes $\tilde{\psi}_i$ are chosen so that they are orthonormal with respect to the $L_2(\mathcal{M})$-inner product. We would like to show that the modes $\tilde{\psi}_i$ are complete in $L_2(\mathcal{M})$. Note that the class of functions $D(\mathcal{M})$, which are bounded and which have compact support on $\mathcal{M}$, is contained in $\{\delta\psi\}_0$. But $D(\mathcal{M})$ is known to be dense in $L_2(\mathcal{M})$ with the $L_2(\mathcal{M})$-norm, and therefore the linear span of the modes $\tilde{\psi}_i$ must be dense in $L_2(\mathcal{M})$. At this point, let us note that the set of functions $L_2(\mathcal{M})$, with the $L_2(\mathcal{M})$-inner product, form a Hilbert space H. It is a standard result that a set of functions $\{\psi_i\}$ is complete in H when the linear span of the functions ψ_i is dense in H, and vice-versa (see, *e.g.*, [33]). This observation implies that the modes $\tilde{\psi}_i$ are complete in $L_2(\mathcal{M})$.

We therefore have two complete and orthonormal sets of functions ψ_i and $\tilde{\psi}_i$ in $L_2(\mathcal{M})$, and an arbitrary field perturbation $\delta\psi \in L_2(\mathcal{M})$ can be expressed in terms of the modes $\tilde{\psi}_i$, *i.e.*,

$$\delta\psi := \psi - \psi_0 = \sum_i \tilde{a}_i \tilde{\psi}_i. \tag{69}$$

It is simple to show that the transformation which expresses one set of basis functions in terms of the other must be orthogonal. Let us now express the measure $\mathrm{d}[\psi]$, given by expression (68), in terms of the new set of modes $\tilde{\psi}_i$. We obtain,

$$\mathrm{d}[\psi] = \prod_i \mu \int \mathrm{d}\tilde{a}_i, \tag{70}$$

where we used that the Jacobian of the transformation relating the coefficients a_i and $\tilde{a}_i$ equals one when the transformation is orthogonal.

One could expect that the path-integral, evaluated with the measures (68) and (70), gives rise to the same result, since all we have done is to express one complete basis of modes in terms of the other. This observation is not correct. Note that when the path-integral (66) is performed with the measure (70), then the source term $J^+[\psi]$ vanishes trivially, since the argument ψ is a linear combination of the modes $\tilde{\psi}_i$, and therefore $\psi \in \{\delta\psi\}_0$. On the contrary, when the path-integral is performed with the measure (68), then ψ is a linear combination of the modes ψ_i, and $J^+[\psi]$ will generally be nonzero, which follows from the observation that $J^+[\psi_i]$ diverges for all ψ_i, as we showed in the previous section.

Let us try to make precise in which sense the expansion of the field in terms of two complete sets of modes (67) and (69) differs. Since both expansions converge to the same limit $\delta\psi$, it follows that the *difference* between the two expansions can only be a configuration with zero $L_2(\mathcal{M})$-norm. When performing the path-integral (66), using the measures (68) and (70) respectively, we are integrating over

paths in $L_2(\mathcal{M})$ which may differ by a zero-norm configuration. These zero-norm configurations are precisely the degrees of freedom which give rise to the nontrivial source term $J^+[\psi]$. In order to show this, let us recall that $J^+[\psi_i]$ diverges for all ψ_i. Since $J^+[\psi]$ is linear in ψ, and $J^+[\tilde{\psi}] = 0$ when $\tilde{\psi}$ is in the linear span of the modes $\tilde{\psi}_i$, it follows that

$$J^+[\psi_i] = J^+[\psi_i - P\psi_i], \tag{71}$$

where $P\psi_i$ denotes the projection of ψ_i onto the basis of modes $\tilde{\psi}_i$, *i.e.*,

$$P\psi_i := \sum_j \langle \tilde{\psi}_j \psi_i \rangle \tilde{\psi}_j. \tag{72}$$

But the modes $\tilde{\psi}_i$ where found to be complete in $L_2(\mathcal{M})$, so that $(1-P)\psi_i$ must have zero $L_2(\mathcal{M})$-norm. The argument of J^+ on the right-hand side of equation (71) has therefore zero $L_2(\mathcal{M})$-norm, and therefore this must be the degree of freedom which causes the divergence of the source term. Since the action functional depends on zero-norm degrees of freedom through the term $J^+[\psi]$, the expression for the correlation function (66) is under-determined. Recall that the same ambiguity was present when we tried to determine the extremal-action configurations in subsection 3.5.2. Although we do not know of a way to resolve this ambiguity, let us consider two different approaches which might work.

First, one can fix the zero-norm degrees of freedom on the basis of a physical or philosophical argument. In practice, this could mean that one sets the source term J^+ equal to zero by restricting the phase-space of the ψ-field to a dense subset of $L_2(\mathcal{M})$ for which J^+ vanishes. In order to make this approach better than just guessing, one needs to establish whether specific restrictions on the phase-space of the ψ-field lead to different predictions, which can be falsified.

As a different approach, one could change the measure on the space of the ψ-field in order to accommodate the zero-norm degrees of freedom. Again, the problem is that there is no clear guideline for doing so, unless one can show that different choices of measure lead to different observable predictions.

It is illustrative to consider a similar ambiguity which occurs in the definition of the path-integral, when one is dealing with fluctuations at infinitesimal rather than infinite length scales. This ambiguity is related to the fact that typical paths which contribute to the path-integral are non-differentiable. Since the class of smooth paths (C^∞) is dense in the class of continuous paths (C^0), the difference between a path in C^0 and the nearest path in C^∞ must have zero $L_2(\mathcal{M})$-norm. As we have seen, the measure (68) does not accommodate these degrees of freedom, and the formal expression is ambiguous on the point of the differentiability of the paths over which we integrate. The action functional is however sensitive to the degree of differentiability of the paths, which is made clear by the fact that the action is generally finite for differentiable paths and infinite for non-differentiable paths. One could try to resolve this ambiguity by simply considering paths in C^∞,

so that the action functional is well defined, but in this case one can show that the field operators in expression (66) commute trivially, and one does not recover quantum physics [19].

Finally, let us note that similar implications hold for other field theories which are described by an action functional which is non-linear in the field variable. In particular, it is well known that the Einstein field equations can be derived by varying an action functional, which is given by

$$S[g_{\mu\nu}] = \frac{1}{16\pi G}\int_{\mathcal{M}} R\sqrt{-g}, \tag{73}$$

where R denotes the Ricci scalar, and we have ommited a possible contribution from matter fields and a cosmological constant. Similar to the case where we considered a scalar field, a contribution of a surface term to the variation of the action does occur. At first-order in the metric perturbation, the contribution of this surface term is given by [34],

$$\delta S[g_{\mu\nu}] = -2\int_{\partial\mathcal{M}} \left(\delta K \ + n^a h^{bc}\delta g_{ab;c}\right) \mathrm{d}\Omega, \tag{74}$$

where δK denotes the variation of the trace of the extrinsic curvature at the boundary $\partial\mathcal{M}$, while h^{bc} and n^a denote the induced three-metric and the normal to the boundary respectively, and $\mathrm{d}\Omega$ denotes the volume element on $\partial\mathcal{M}$. The first term on the right-hand side of equation (74) can be canceled by adding a surface integral of two times the extrinsic curvature K to the action functional (73) (see also the discussion in section 3). The second term on the right-hand side of equation (74) vanishes when it is evaluated according to a classical variational approach where we set δg_{ab} equal to zero at the bounday $\partial\mathcal{M}$, but this term could be of interest in cosmological situations when we do not require that perturbations vanish outside a finite volume.

3.7 Conclusion

We revisited the variational principle in a cosmological context. Starting from the path-integral formulation of quantum physics, we argued that there is a correspondence between classical physics and extremal action fields. The phase-space in which extremality of the action is considered, is not constrained in quantum physics, and we showed that there can be a non-trivial contribution arising from surface terms. We made this problem explicit by considering a scalar field in a perturbed open FLRW spacetime. In the case of an open FLRW spacetime with a Gaussian spectrum of subcurvature perturbations, we found no non-trivial correction to the classical equation of motion. In the case where supercurvature perturbations are present, extremality of the action could only be defined after adopting additional restrictions on the phase-space of the scalar field, but the corresponding equations of motion are ambiguous since they depend on how one restricts the

phase-space of the field. We showed that the restricted phase-spaces which yield different physical results, differ by perturbations with vanishing L_2-norm. This ambiguity is present both at a classical level and a quantum level. We briefly discussed a possible strategy to resolve the ambiguity which is due to perturbations with vanishing L_2-norm.

References

[1] M. MacCallum, in *General relativity, an Einstein centenary survey*, edited by S. W. Hawking and W. Israel (Cambridge University Press, Cambridge, 1979).

[2] G. E. Sneddon, J. Phys. **A 9**, 229 (1976).

[3] M. MacCallum and A. Taub, Commun. Math. Phys. **25**, 173 (1972).

[4] M. Ryan, J. Math. Phys. **15**, 812 (1974).

[5] M. Bander and C. Itzykson, Rev. Mod. Phys. **38**, 346 (1966).

[6] D. H. Lyth and A. Woszczyna, Phys. Rev. **D 52**, 3338 (1995).

[7] J. D. Cohn and D. I. Kaiser, gr-qc/9803073.

[8] D. I. Kaiser, astro-ph/9608025.

[9] B. Ratra, Phys. Rev. **D 50**, 5252. (1994).

[10] B. Ratra and P. J. E. Peebles, Astrophys. J. Lett. **432**, L5 (1994).

[11] M. Bucher, A. S. Goldhaber and N. Turok, Phys. Rev. **D 52**, 3314 (1995).

[12] M. Sasaki *et al.*, Phys. Lett. B **317**, 510 (1993).

[13] A. Stebbins and R. R. Caldwell, Phys. Rev. **D 52**, 3248 (1995).

[14] B. Ratra and P. J. E. Peebles, Phys. Rev. **D 52**, 1837 (1995).

[15] K. Yamamoto *et al.* Phys. Rev. **D 51**, 2968 (1995).

[16] Misao Sasaki *et al.*, Phys. Rev. **D 51** 2979 (1995).

[17] E. Noether, *Invariante Variations Probleme*, Nachr. Ges. Gottingen, 235- 257 (1918).

[18] D. Bleecker, *Gauge Theory and Variational Principles*, Addison-Wesley (1981)..

[19] R. J. Rivers, in *Path integral methods in quantum field theory*, edited by P. V. Landshoff *et al.* (Cambridge University Press, Cambridge, 1987).

[20] J. B. Hartle, in *Quantum cosmology and baby universes*, edited by S. Coleman *et al.* (World Scientific, Singapure, 1991).

[21] T. Regge and C. Teitelboim, Ann. Phys., **88**, 286-318.

[22] S. W. Hawking, in *General relativity, an Einstein centenary survey*, edited by S. W. Hawking and W. Israel (Cambridge University Press, Cambridge, 1979).

[23] J. B. Hartle and S. W. Hawking, Phys. Rev. **D 28**, 2960 (1983).

[24] R. Fabbri, I. Guidi and V. Natale, Astrophys. J. **257**, 17 (1982).

[25] J. Garcia-Bellido, Phys. Rev. **D 55**, 4596 (1995).

[26] R. J. Adler, *The Geometry of Random Fields* (Wiley, Chichester, 1981).

[27] S. Karin and H. M. Taylor, *A first course on stochastic processes*, (Academic Press, New York, 1975).

[28] J. M. Bardeen *et al.*, Astrophys. J. **304**, 15 (1986).

[29] A. M. Yaglom, in *Proceedings of the Fourth Berkeley Symposium Volume II*, edited by J. Neyman (University of California Press, Berkeley, 1961).

[30] L. F. Abbott and R. K. Schaefer, Astrophys. J. **308**, 546 (1986).

[31] F. G. Friedlander, *Introduction to the theory of distributions* (Cambridge University Press, Cambridge, 1982).

[32] S. W. Hawking, in *General relativity, an Einstein centenary survey*, edited by S. W. Hawking and W. Israel (Cambridge University Press, Cambridge, 1979).

[33] R. F. Streater and A. S. Wightman, *PCT, Spin and Statistics, and All That* (Benjamin, New York, 1964).

[34] R. M. Wald, *General Relativity* (University of Chicago Press, Chicago, 1984).

4 Open inflation

Abstract

We discuss the idea of spatially flat inflation and spatially open (bubble) inflation in a curved spacetime with a scalar field. Although the bubble inflation scenario explains the creation of an open Friedmann-Lemaître-Robertson-Walker (FLRW) universe, this process can only occur for a restricted class of potentials. We discuss the dynamics of a spherically symmetric bubble-spacetime in some detail, and we argue that the often used thin-wall approximation is not consistent with the stress-energy of a decaying scalar field. As an alternative, we derive a simplified set of equations which describes the exact dynamics of a spherically symmetric bubble-spacetime (without solving these equations). We then consider the possibility that a negatively curved spacetime is generated through inflation with a general potential and random initial conditions. It is shown that the spacelike hypersurfaces on which the scalar field is constant start their evolution with a singular negative spatial curvature and a vanishing kinetic contribution to the stress-energy. This result shows that negative spatial curvature can be generated naturally during inflation, without restricting the type of potential, but given the assumption of a continuous change of the spatial volume volume element in terms of a comoving coordinate system. Further, we comment on the possibility of a discontinuous evolution of the metric in terms of a comoving coordinate system.

4.1 Introduction

The idea of inflation, which was first proposed by Guth in [1], provides a simple explanation for both the flatness of the observed universe, as well as the observed spectrum of perturbations at the surface of last scattering. The key assumption of the inflation scenario is that one assumes an equation of state for which the pressure p is of the order of minus the energy density ρ. As is well known, the energy density varies slowly as a function of time for this equation of state, and hence the constraint equations imply that the volume expansion is nearly constant in time. A nearly constant expansion implies that the physical volume of a spatial hypersurface tends to grow nearly exponentially in time. The effect of this expansion is that inhomogeneities are stretched, while the curvature of spatial hypersurfaces tends to zero. Therefore an inflating spacetime becomes spatially very flat and homogeneous at late times, when considered at a fixed physical length scale. Since inflation eliminates inhomogeneities and nonzero spatial curvature by the same mechanism, one expects that a universe which has acquired spatial homogeneity through inflation must also be spatially very flat, with deviations from

spatial homogeneity and flatness of the same order. Hence, it seems to be puzzling that inhomogeneities in the observed universe appear to be very small, of the order of 10^{-5} at the surface of last scattering [2], while observations indicate that the universe is negatively curved, with a radius of spatial curvature which is of the same order as the Hubble radius (see, *e.g.*, [3]).

In order to reconcile the observed large-scale negative spatial curvature of the universe with the large degree of smoothness and isotropy, two important ideas have been developed. Firstly, there is the 'bubble' scenario, which describes the emergence of a spatially homogeneous and isotropic negatively curved spacetime inside a 'bubble', which is a bounded region in an exponentially expanding spacetime. The exponential expansion of the spacetime which surrounds the bubble is generated by a cosmological constant type of stress-energy tensor. This type of stress-energy is thought to be generated by a scalar field which is trapped in a local minimum of its potential, and this state is usually referred to as a 'false-vacuum state'. Due to the exponential expansion, a patch of the spacetime in which the bubble is embedded is very accurately described by a patch of De Sitter spacetime. As is well known, De Sitter spacetime allows for a choice of coordinates in which the metric on a section of the spacetime takes the form of a negative spatially curved (open) FLRW metric. In order to generate a spatially open FLRW metric, which is radiation or matter dominated at late times, the scalar field must decay from the false-vacuum state to the true-vacuum state, and the decay of the scalar field must happen synchronously in terms of a coordinate system in which the metric has the open FLRW form. Whether this condition is satisfied depends on the dynamics of the false-vacuum state, as well as the mechanism which initiates the decay of the scalar field.

An alternative mechanism which describes the creation of an open FLRW universe has recently been proposed by Hawking and Turok in [4]. According to their proposal, the spacetime which surrounds the bubble with the open FLRW geometry is replaced by a spatially closed geometry which is singular at a point where the scalar field diverges. Given certain a priori assumptions, it is conceivable that this geometrical structure originates through quantum processes in an abstract space of geometries and field configurations. It is, however, not our aim to discuss this idea in detail, and in the following we will concentrate on the bubble scenario, and we present an alternative mechanism which describes the generation of a negative spatially curved spacetime through inflation, without assuming the existence of a false-vacuum state.

This chapter is structured as follows. In section 4.2 we will discuss spatially flat inflation in a general inhomogeneous spacetime. In section 4.3 we discuss the generation of an open FLRW spacetime inside a bubble which is embedded in a De Sitter geometry. We also derive two coupled differential equations which describe the dynamics of a general $O(1,3)$ symmetric bubble-spacetime. In subsection 4.4 we discuss the thin-wall description of bubble-dynamics, as well as various reasons

why this method can be questioned when it is applied to describe a realistic bubble. In subsection 4.5 we derive a simplified set of equations which describes the exact dynamics of an arbitrary spherically symmetric bubble-spacetime.

In section 4.6 we discuss the possibility that a negatively curved and approximately homogeneous spacetime is generated by inflation with random initial conditions and without assuming the existence of a false-vacuum state. We study the dynamics of the geometry and the scalar field in terms of comoving coordinates in subsection 4.6.1. In subsection 4.6.2 it is shown that the constant scalar field hypersurfaces start their spacelike evolution with a singular negative spatial curvature, or the induced metric of these hypersurfaces must evolve discontinuously as a function of time. In subsection 4.6.3 we show that the kinetic part of the stress-energy must vanish at the points where the constant scalar field hypersurfaces make a transition from being timelike to being spacelike, and we show that there is only one choice of an arrow of time in our spacetime which is physically acceptable. In subsection 4.6.4 we discuss an alternative mechanism which could explain the generation of negative spatial curvature, as well as approximate spatial homogeneity and isotropy.

4.2 Inflation

In this section we will discuss some of the more technical aspects which underly the idea of inflation. Although most of the ideas which we consider in this section are well known and published, the approach which we follow differs from most other treatments of this subject on the point that we do *not* assume from the start that the spacetime is (nearly) spatially homogeneous and isotropic. This allows us to investigate situations in which the spacetime is manifestly inhomogeneous. In particular, the approach which we follow provides a useful introduction to the following sections, where we consider the possibility that a negatively curved FLRW spacetime is generated during inflation without assuming that the scalar field is initially in a false-vacuum state.

We consider the case where inflation is driven by a real Klein-Gordon field ϕ with a potential term $V(\phi)$ in the Lagrangian. The Lagrangian density of a real Klein-Gordon field is given by,

$$\mathcal{L} = -\frac{1}{2}\sqrt{-g}(\partial_\mu\phi\partial^\mu\phi + 2V(\phi)) \tag{75}$$

where $V(\phi)$ is a function of ϕ which is bounded from below. The classical equation of motion for the scalar field is given by,

$$\frac{1}{\sqrt{-g}}\partial_\mu\sqrt{-g}g^{\mu\nu}\partial_\nu\phi - V_{,\phi} = 0, \tag{76}$$

where $V_{,\phi} := \frac{\partial}{\partial\phi}V(\phi)$, and the stress-energy tensor of the scalar field follows by functionally differentiating the Lagrangian density (75) with respect to the metric

$g^{\mu\nu}$ (see, *e.g.*, [5]),

$$T_{\mu\nu} := -\frac{2}{\sqrt{-g}}\frac{\delta\mathcal{L}}{\delta g^{\mu\nu}} \tag{77}$$

$$= \partial_\mu\phi\partial_\nu\phi \; - \frac{1}{2}g_{\mu\nu}(\partial^\mu\phi\partial_\mu\phi + 2V(\phi)).$$

We define n_μ to be the unit vectors normal to the hypersurfaces on which ϕ is constant, *i.e.*,

$$n_\mu := (|\partial^\mu\phi\partial_\mu\phi|\,)^{-\frac{1}{2}}\partial_\mu\phi \tag{78}$$

where we assume that $\partial^\mu\phi\partial_\mu\phi \neq 0$ (the case where $\partial^\mu\phi\partial_\mu\phi$ vanishes will be dealt with in section 4.6). We may use the definition (78) to write the scalar field stress-energy tensor (77) in the perfect fluid form,

$$T_{\mu\nu} = \epsilon(\rho + p)n_\mu n_\nu + pg_{\mu\nu}, \tag{79}$$

where, from now on, ϵ equals $-1(1)$ in the case where n_μ is timelike (spacelike). We should note that the $\pm$ sign in expression (79) is no more than a convention, and one could absorb a minus sign in the definitions of ρ and p. Combining expression (77) and (79) yields the following expressions for the energy density ρ and the pressure p,

$$\rho = -\frac{1}{2}\partial^\mu\phi\partial_\mu\phi + V(\phi), \tag{80}$$

and

$$p = -\frac{1}{2}\partial^\mu\phi\partial_\mu\phi - V(\phi). \tag{81}$$

Let us now define *comoving* coordinates by the condition that hypersurfaces of constant comoving time parameter are orthogonal to n_μ, and the lines of constant spatial coordinates are integral curves of n_μ (see also the appendix 4.B). Note that the hypersurfaces of constant comoving time coincide with the hypersurfaces on which the ϕ-field is constant, and these hypersurfaces are spacelike (timelike) in the case where n_μ is timelike (spacelike). It is now convenient to write the metric in the form

$$g_{\mu\nu} = \epsilon n_\mu n_\nu + h_{\mu\nu}, \tag{82}$$

where $h_{\mu\nu} := g_{\mu\nu} - \epsilon n_\mu n_\nu$ is the induced metric on the constant-ϕ hypersurfaces. Inserting the expression for the metric, (82), into the field equation, (76), we obtain,

$$0 = \ddot{\phi} + n^\mu_{;\mu}\dot{\phi} - \epsilon V_{,\phi}\ , \tag{83}$$

where a dot denotes the Lie-derivative with respect to the vector field n^μ, and a semicolon denotes covariant differentiation.

Multiplying the field equation (83) by $\dot{\phi}$, and using the definitions (80) and (81), we recover an energy conservation equation,

$$0 = \dot{\rho} + n^\mu_{;\mu}(\rho + p). \tag{84}$$

The divergence of the normals n^μ which appears in equation (84) can be shown to be equal to the trace of the extrinsic curvature of the constant-ϕ hypersurfaces. In the appendix 4.B it is shown that

$$n^\mu_{;\mu} = K := h^{\mu\nu} K_{\mu\nu}, \tag{85}$$

where

$$K_{\mu\nu} := \frac{1}{2} \mathcal{L}_n h_{\mu\nu} \tag{86}$$

is referred to as the extrinsic curvature of the constant-ϕ hypersurfaces. A typical equation of state for which inflation can occur corresponds to the situation where

$$\rho \gg \rho + p, \tag{87}$$

and $\rho > 0$. According to equations (80) and (81), a scalar field with a potential term generates an equation state of the form (87) provided that the squared gradient of the ϕ-field is small compared to the potential $V(\phi)$. When we assume that condition (87) is satisfied, then the conservation equation (84) implies that,

$$\left| \frac{\dot{\rho}}{K\rho} \right| \ll 1, \tag{88}$$

which means that the energy density ρ changes by a small fraction over a time scale of the order of the expansion time, which we define as K^{-1}.

The approximate constancy of the energy density ρ implies, under certain conditions, approximate constancy of the expansion K at a time scale of the order of K^{-1}. The relation between the energy density and the expansion is provided by the constraint equation,

$$0 = R^{(3)} + \frac{2}{3} K^2 - \sigma_{\mu\nu} \sigma^{\mu\nu} - 2\kappa\rho, \tag{89}$$

where $\sigma_{\mu\nu} := K_{\mu\nu} - \frac{1}{3} K h_{\mu\nu}$ is the shear of the normals n^μ, $R^{(3)}$ denotes the spatial curvature scalar of the three-dimensional hypersurfaces which are orthogonal to n^μ, and $\kappa := 8\pi G$.

It should be noted that the constraint equation (89) holds for any foliation of the spacetime in terms of a collection of hypersurfaces for which the normals are not null (see, *e.g.*, the derivation of this equation in [5]). In the following, we will consider the case where the constant-ϕ hypersurfaces are spacelike.

Let us observe that the curvature term in equation (89) may have any sign, while the squared shear is positive definite. When the spatial curvature is positive and of the order of the sum of the last two terms in equation (89), then K may vanish. Since we have not derived any constraints on the three-curvature of the constant-ϕ hypersurfaces, there is not much we can say rigorously about the relation between the energy density ρ and the expansion K. However, it seems natural to assume

that the spatial curvature $R^{(3)}$ approaches zero as a function of time, provided that the spacetime is expanding. Under this assumption, equation (89) yields an approximate expression for the expansion K,

$$K^2 \gtrsim \frac{3\kappa\rho}{2}, \tag{90}$$

where we used that the squared shear in equation (89) is positive definite, and hence this term contributes positively to the value of K^2. Since ρ was shown to be approximately constant over a time scale of the order of the expansion time K^{-1} for an equation of state which satisfies condition (87), it follows that the expansion K is approximately constant as a function of time. Note also that equation (90) does not determine the sign of the expansion K. The case where the expansion K is negative corresponds to the time reverse of the case where K is positive, and it seems therefore natural to choose an arrow of time such that K is positive.

It is shown in the appendix that K equals the Lie-derivative with respect to n^μ of a physical volume element which is associated with a fixed coordinate volume in the comoving gauge, *i.e.*,

$$K = \dot{(\ln \sqrt{h})}, \tag{91}$$

where $h := \det(h_{\mu\nu})$. Indeed, approximate constancy of K over a time scale of the order of K^{-1} implies approximate *exponential* growth of h, *i.e.*,

$$\frac{\sqrt{h(t)}}{\sqrt{h(0)}} = \exp[\int_0^{x^0=t} \mathrm{d}x^0 N\ K(x^0)] \approx \exp[K(0)Nt], \tag{92}$$

where $N = \sqrt{g_{00}}$, and x^0 denotes the comoving time variable. Expression (92) does not tell us whether the spatial metric of the comoving hypersurfaces expands nearly isotropically, in the sense that the anisotropic part of the extrinsic curvature, $\sigma_{\mu\nu}$, can be neglected as compared to the isotropic part, which equals $\frac{1}{3}Kh_{\mu\nu}$. Although it seems plausible that the shear $\sigma_{\mu\nu}$ can be neglected as compared to the isotropic expansion $\frac{1}{3}Kh_{\mu\nu}$ at late times after inflation has started, this has only been proven in the case where the spacetime is spatially homogeneous [6]. Indeed, efforts have been made to generalize this result to general inhomogeneous and anisotropic spacetimes (see, *e.g.*, [7, 8]), without rigorous results. Let us now assume that the expansion is isotropic, and that the geometry of the comoving spatial hypersurfaces is smooth at some comoving length scale, then it follows that a patch of the spacetime approaches a spatially flat FLRW geometry when considered at a fixed physical length scale. Hence, one expects that a patch of a nearly exponentially expanding spacetime will locally be very similar to a patch of De Sitter spacetime.

4.3 Bubble inflation

In the previous section we showed that a cosmological constant type of equation of state, which is generated by a scalar field with a potential term, causes an approximately exponential expansion of the spatial volume of the constant-ϕ hypersurfaces. The significance of the constant-ϕ hypersurfaces in the inflation scenario stems from the fact that these hypersurfaces provide a physical interpretation to the 3 + 1 splitting of the four-dimensional spacetime in terms of a collection of spatial hypersurfaces. Namely, one expects that the spatial hypersurfaces on which the scalar field is constant coincide approximately with hypersurfaces of constant energy density. At late times, after a period of inflation, these hypersurfaces of constant energy density determine the shape of the surface of last scattering, on which we base our perceptions of the spatial structure of the early universe. An interesting solution of the field equation is the false-vacuum solution, which is characterized by the vanishing of the gradients of the scalar field, while the potential term which contributes to the stress-energy tensor (77) does not vanish. It follows from the field equation (83) that the gradient of the scalar field can only vanish everywhere when the potential is locally flat, *i.e.*, $V_{,\phi} = 0$. Further, one expects that stability of the constant field solutions requires that $V_{,\phi,\phi}$ is positive, which means that the field is at a local minimum of the potential. It follows from the expression for the scalar field stress-energy that the false-vacuum stress-energy is described by a contribution to the cosmological constant which is equal to $V(\phi)$. In the case where the cosmological constant is positive, we know that there is an exact solution of Einstein's equation, which is De Sitter spacetime. The importance of the false-vacuum state in the bubble-inflation scenario is related to the fact that the decay of a false-vacuum can naturally give rise to an open FLRW universe, which is contained in a bounded region, *i.e.*, a 'bubble', which is embedded in a De Sitter spacetime. In order to clarify the bubble-scenario, we will first discuss the De Sitter geometry in some more detail, and then we consider the deformation of the De Sitter geometry which occurs when there is a spherically symmetric decay of the false-vacuum state. A convenient representation of the De Sitter geometry is obtained by embedding this geometry in a 1+4 dimensional Minkowski spacetime, *i.e.*, we consider the collection of points which satisfy the condition

$$(x^0)^2 - \sum_{i=0}^{4} (x^i)^2 = -H^{-2}, \tag{93}$$

where x^0 and $x^i (i \in \{1 \cdots 4\})$ denote the time and space coordinates respectively in a 1+4 dimensional Minkowski spacetime with signature $(-1, 1, 1, 1, 1)$, and H can be interpreted as the Hubble parameter in a spatially flat coordinate system, which we will introduce subsequently. In the following, we will refer to the hyperboloid which is defined by condition (93) as $\mathcal{H}$, and we choose our unit of length such that the Hubble constant H equals one. The group of isometries on $\mathcal{H}$ corresponds to the group of isometries in 1+4 dimensional Minkowski spacetime, restricted by

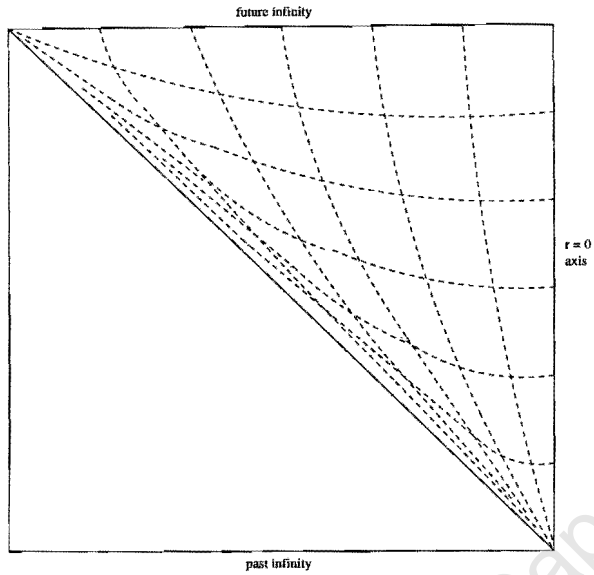

Figure 4: Conformal diagram of De Sitter spacetime. Each point in the diagram represents a two-sphere. Spacelike and timelike dashed lines represent hypersurfaces of constant time and comoving spatial coordinate in the flat coordinate system (95).

the condition that points are mapped from $\mathcal{H}$ to $\mathcal{H}$. This group of isometries corresponds to $O(1,4)$, which is also called the De Sitter group.

As is well known, one can choose coordinates on $\mathcal{H}$ such that the metric on a section of $\mathcal{H}$ takes the spatially flat FLRW form. Let us introduce a new set of coordinates $\{t, x, y, z\}$ by,

$$t = \ln(x^0 + x^1) \ , \quad x = \frac{x^2}{x^0 + x^1} \ , \quad y = \frac{x^3}{x^0 + x^1} \ , \quad z = \frac{x^4}{x^0 + x^1} \, . \tag{94}$$

One can show that the coordinates $\{t, x^i\}$ yield a spatially flat FLRW metric in the region of $\mathcal{H}$ for which $z^0 + z^4 > 0$ (see, *e.g.*, [9]),

$$\mathrm{d}s^2 = -\mathrm{d}t^2 + e^{2t}[\mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2] \, . \tag{95}$$

Further, the region in $\mathcal{H}$ for which $z^0 + z^4 < 0$ can be coordinatized by making the substitution $z^\mu \to -z^\mu (\mu \in \{0 \cdots 4\})$ in expression (94), and the metric in the region of $\mathcal{H}$ for which $z^0 + z^4 < 0$ appears to be identical to the metric (95).

Apart from the spatially flat coordinatization, (95), there exists a coordinate system for which the metric in four different regions of $\mathcal{H}$ takes the spatially open FLRW form. Let us therefore define the coordinates $\{\tau, \chi, \theta, \phi\}$ by

$$x^0 = \sinh\tau \cosh\chi \ , \quad x^1 = \cosh\tau \ , \quad x^2 = \sinh\tau \sinh\chi \sin\theta \cos\phi,$$

$$x^3 = \sinh\tau \sinh\chi \sin\theta \sin\phi \ , \quad x^4 = \sinh\tau \sinh\chi \cos\theta, \tag{96}$$

and one finds that the metric on a section of $\mathcal{H}$, which we call region I, takes the form,

$$ds^2 = -d\tau^2 + \sinh^2\tau[d\chi^2 + \sinh^2\chi d\Omega^2], \tag{97}$$

where $\tau \in \mathbf{R}, \chi \in \mathbf{R}^+, \phi \in [0, 2\pi], \theta \in [0, \pi]$, and $d\Omega^2 = d\theta^2 + \sin^2\theta d\phi^2$ is the surface element on a unit two-sphere.

Note that the metric (97) describes *two* open FLRW geometries, since τ takes values in $\mathbf{R}^+$ as well as $\mathbf{R}^-$. There is another section of $\mathcal{H}$ which can be coordinatized such that the metric takes the form (97), and the explicit coordinate transformation is obtained by making the substitution $x^\mu \to -x^\mu (\mu = 0 \ldots 4)$ in expression (96). Note also that the origin of the spatially open sections, which corresponds to $\tau = 0$ for a constant value of χ in the metric (97), can be chosen to be any point on the De Sitter hyperboloid $\mathcal{H}$. Hence, we find that a different spatially open coordinate system is associated with every element of the De Sitter group which does not leave invariant the position of the origin of the spatially open coordinate system.

The section of $\mathcal{H}$ which is not covered by the coordinates $\{\tau, \chi, \theta, \phi\}$, can be covered by a different set of coordinates $\{\tilde{\tau}, \tilde{\chi}, \tilde{\theta}, \tilde{\phi}\}$, which we define by,

$$x^0 = \sinh\tilde{\tau} \sin\tilde{\chi} \ , \quad x^1 = \cos\tilde{\theta} \ , \quad x^2 = \cosh\tilde{\tau} \sin\tilde{\chi} \cos\tilde{\theta}$$

$$x^3 = \cosh\tilde{\tau} \sin\tilde{\chi} \sin\tilde{\theta} \cos\tilde{\phi} \ , \quad x^4 = \cosh\tilde{\tau} \sin\tilde{\chi} \sin\tilde{\theta} \sin\tilde{\phi},$$

and the metric on the section of $\mathcal{H}$ which is covered by the coordinates $\{\tilde{\tau}, \tilde{\chi}, \tilde{\theta}, \tilde{\phi}\}$, takes the form,

$$ds^2 = d\tilde{\chi}^2 + \sin^2\tilde{\chi}\,[-d\tilde{\tau}^2 + \cosh^2\tilde{\tau} d\Omega^2], \tag{98}$$

where $\tilde{\tau} \in \mathbf{R}, \tilde{\chi}, \tilde{\phi} \in [0, 2\pi], \tilde{\theta} \in [0, \pi]$. It is clear that the line element (97) does not describe a realistic open FLRW geometry, since the stress-energy tensor in $\mathcal{H}$ is proportional to the metric, which is invariant under the De Sitter group. Hence, there is no physical criterion which singles out a preferred coordinate system for which the metric takes the spatially open FLRW form. In order to obtain a physically realistic open FLRW geometry, the scalar field, which determines the stress-energy, must vary as a function of the time coordinate τ, while it is independent of χ, θ, and ϕ. Let us denote a spatial hypersurface for which the time coordinate τ in region I is constant, by $\Sigma(\tau)$. Similarly, we denote a timelike hypersurface in region II, for which the coordinate $\tilde{\chi}$ is constant, by $\tilde{\Sigma}(\tilde{\chi})$. The group of isometries in a realistic open FLRW universe, where ϕ varies as a function of τ, must be a subgroup of the De Sitter group which maps points from $\Sigma(\tau)$ to $\Sigma(\tau)$. The subgroup of the De Sitter group which satisfies this condition corresponds to $O(1,3)$, and it can be visualized as the group of spatial translations, rotations, and reflections with respect to a point at the hypersurfaces $\Sigma(\tau)$. Similarly, one can show that the subgroup of the De Sitter group which maps points from $\tilde{\Sigma}(\tilde{\chi})$ to $\tilde{\Sigma}(\tilde{\chi})$ in region II corresponds to $O(1,3)$.

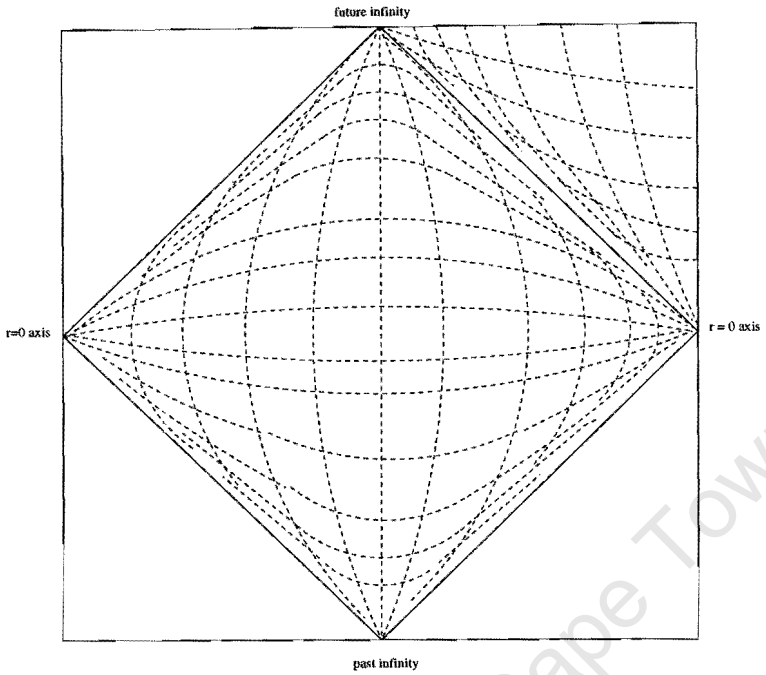

Figure 5: Conformal diagram of De Sitter spacetime. In the four triangular sections which we call region I, the metric takes the open FLRW form (97), while the central square, which we call region II, is described by the metric (98).

The reason for studying an $O(1,3)$ invariant solution of the field equation is that this provides some insight in the dynamics of the scalar field in region II, where the constant-ϕ hypersurfaces are spacelike. When studying restricted classes of solutions of the field equations, one should note that solving the field equations and then selecting a solution, which satisfies the required symmetry condition, may not lead to the same result as first imposing symmetry conditions and then solving the field equation for the remaining degrees of freedom. Clearly, the first approach is correct in the sense that it yields solutions of the unrestricted field equations which possess the desired symmetry condition, if any solutions exist. However, the second approach has the advantage that it is relatively simple to solve the field equations for the degrees of freedom which remain after imposing a symmetry condition, but it is clear that one will have to verify whether solutions of this type satisfy the unrestricted field equations.

The requirement that the scalar field ϕ is invariant under the $O(1,3)$ subgroup of the De Sitter group which maps points from $\Sigma(\tau)$ to $\Sigma(\tau)$ implies that ϕ can only be a function of the coordinate τ in region I. Similarly, it follows that ϕ can only be a function of the coordinate $\tilde{\chi}$ in region II. The line element which describes an

$O(1,3)$ invariant deformation of the De Sitter geometry is given by

$$\mathrm{d}s^2 = -\mathrm{d}\tau^2 + \alpha^2(\tau)\sinh^2\tau\ [\mathrm{d}\chi^2 + \sinh^2\chi \mathrm{d}\Omega] \tag{99}$$

in region I, and

$$\mathrm{d}s^2 = \mathrm{d}\tilde{\chi}^2 + \alpha^2(\tilde{\chi})\sin^2\tilde{\chi}\ [-\mathrm{d}\tilde{\tau}^2 + \cosh^2\tilde{\tau}\mathrm{d}\Omega^2] \tag{100}$$

in region II. The De Sitter line element is recovered as a special case of the line elements (99) and (100) when we set α equal to one for all τ and $\tilde{\chi}$.

The scalar field equation of motion in region I is given by expression (83), where $\epsilon = -1$ and dot denotes $\partial/\partial\tau$. Further, the volume expansion which enters the equation of motion is easily calculated by using expression (97) for the metric in region I. We obtain,

$$K = \frac{3}{2}[\frac{\dot{\alpha}}{\alpha} + \ \mathrm{cotanh}\tau]. \tag{101}$$

Similarly, in region II the scalar field equation of motion is given by expression (83), where $\epsilon = 1$, a dot denotes $\partial/\partial\tilde{\chi}$, and

$$K = \frac{3}{2}\ [\frac{\dot{\alpha}}{\alpha} + \mathrm{cotan}\tilde{\chi}]. \tag{102}$$

Note that the field equation (83), when applied to region I as well as region II, describes the dynamics of a field in an expanding hyperbolic geometry, which starts with a (coordinate) singularity at the point where $\tau = \tilde{\chi} = 0$. In spite of this similarity, the interpretation of the field equation in region I and region II differs in the sense that τ is a time coordinate while $\tilde{\chi}$ is a spatial coordinate. Hence, it does not seem correct to interpret the field equation (83), when applied to region II, as an *evolution* equation for the field given certain initial conditions. Instead, it seems more appropriate to consider the field equation in region II as an *a priori* consistency requirement which applies to $O(1,3)$ invariant solutions of the field equation in region II.

It is also of interest that the derivative of the potential enters with opposite signs in the equations of motion in region I and region II respectively. Hence, an $O(1,3)$ invariant solution of the field equation may oscillate about a local *maximum* of the potential as a function of $\tilde{\chi}$ in region II. It is also clear that the evolution of the scalar field towards the true minimum of the potential will be concentrated in region I.

The general relativistic constraint equation (89) takes the form of a second-order differential equation for the metric variable $\alpha(\tilde{\chi})$ when it is applied to the $O(1,3)$ invariant geometry in region II. The explicit expression for the constraint equation (89) in terms of the metric variable α is easily obtained. Note that the volume expansion K was already calculated, and it is given by expression (102). Further, the three-curvature of the constant-$\tilde{\chi}$ hypersurfaces follows when we use

that the induced metric on the hypersurfaces $\tilde{\Sigma}(\tilde{\chi})$ equals the square of a scale factor $\alpha \sin \tilde{\chi}$ times the metric of a $2+1$ dimensional unit-hyperboloid. The $2+1$ dimensional unit hyperboloid has a Ricci scalar equal to -6, and hence,

$$R^{(3)} = -\frac{6}{(\alpha \sin \tilde{\chi})^2} \,. \tag{103}$$

The energy density ρ which appears in equation (89) is given by expression (77), and hence it follows that ρ is *negative* in region II in the case where the potential $V(\phi)$ is positive. By substituting expressions (80), (102), and (103) for ρ, K and $R^{(3)}$, respectively, into the constraint equation (89), we obtain a first-order nonlinear differential equation in α and ϕ, namely,

$$0 = \frac{3}{2}[\frac{\dot{\alpha}}{\alpha} + \text{cotan}\tilde{\chi}]^2 \; - 6(\alpha \sin \tilde{\chi})^{-2} + \kappa(\dot{\phi}^2/2 - V(\phi)), \tag{104}$$

where a dot denotes $\partial/\partial\tilde{\chi}$. Equation (104), in combination with the field equation (83), determines the scalar field and the geometry of an $O(1,3)$ invariant bubble-spacetime. It would be of interest to integrate these two coupled second-order differential equations numerically for a realistic class of potentials $V(\phi)$. In particular, this would provide a test for approximations (such as the thin-wall method) on which most perceptions of the dynamics of bubbles is based.

4.4 The thin-wall method

The thin-wall method is often used to model the dynamics of a spacetime in which a false-vacuum region is separated from a true-vacuum region by a boundary (which we call a 'bubble-wall') of which the thickness is much smaller than the Hubble radius. Whether or not a realistic bubble-wall satisfies the thin-wall criterion can be expected to depend on the shape of the potential. One expects that a steeper slope in between the two minima of the potential generates a faster transition from the false-vacuum state to the true-vacuum state. Hence, the thin-wall approximation is applicable to the case where the local and the true minimum of the potential are separated by a sufficiently steep slope. In the case where the bubble-wall is sufficiently thin, one could try to idealize the bubble-wall by a 2+1 dimensional hypersurface which separates two exact solutions of Einstein's equations, in which the stress-energy tensor has the false-vacuum and the true-vacuum form respectively. The regularity of the metric requires that the two solutions have the same induced metric at the 2+1 dimensional hypersurface where these solutions are identified. Further, when we assume that the bubble-geometry is a solution of Einstein's equation, then the stress-energy tensor of the bubble-wall can be determined from the metric at the bubble-wall by using the Gauss-Codacci relations [10].

A specific idealized geometry to which the thin-wall method can be applied consists of a patch of Minkowski or Schwarzschild spacetime, which is connected

to an exterior region which is described by the De Sitter metric. The boundary between the two geometries forms a $2+1$ dimensional hypersurface, which we call the bubble-wall. For simplicity, we assume that the metric possesses spatial spherical symmetry, which we define by the condition that there exists a choice of coordinates such that the metric is invariant under a group of spatial rotations. Since this geometry is often used as a model for a realistic spacetime where a false-vacuum decays to a true-vacuum [1, 3, 11, 12, 13], we will consider this method in some more detail. In the case where the bubble-wall is spatially spherically symmetric, its dynamics is fully determined by its surface radius $r = r(t)$, which we define as the square root of $1/4\pi$ times the surface of the bubble-wall at a constant time t.

Let us now define η to be a coordinate which equals zero only at the bubble-wall, such that the vector $\partial/\partial\eta$ is the normal vector to the bubble-wall. Without loss of generality we can choose our coordinates such that

$$g^{\eta\eta} = g_{\eta\eta} = 1, \tag{105}$$

and $g^{\eta i} = g_{\eta i} = 0$ follows since $\partial/\partial\eta$ is normal to the bubble-wall. Further, let x^0 denote a time coordinate, which we can choose such that the metric satisfies the condition,

$$g_{00} = -1 \text{ and } g_{0i} = 0, \tag{106}$$

at the bubble-wall, and we let x^2 and x^3 denote the two angular coordinates on a two-sphere of constant time x^0.

By construction, the stress-energy tensor in the spacetime, which consists of an exterior region of De Sitter spacetime and an interior region of Minkowski spacetime, has the form

$$T^{\mu\nu} = \theta(\eta)\Lambda g^{\mu\nu} + S^{\mu\nu}\delta(\eta), \tag{107}$$

where $\theta(\eta) := 1$ for $\eta > 0$, which corresponds to the false-vacuum region, and $\theta(\eta) := 0$ for $\eta < 0$, which corresponds to the Minkowski region. We will determine the tensor $S^{\mu\nu}$ through Einstein's equations.

Using the Gauss-Codacci formalism, the geometrical part of Einstein's equations can be expressed in terms of the intrinsic and the extrinsic curvature on an arbitrary $2+1$ dimensional hypersurface. These equations have been evaluated in [1, 10], and they have the form,

$$K^2 - K_{ij}K^{ij} - {}^{(3)}R = 2\kappa T^{\eta}_{\eta}, \tag{108}$$

$$K^j_{i\ |j} - K_{|i} = \kappa T^{\eta}_i, \tag{109}$$

$$\begin{aligned} {}^{(3)}R^i_j - \frac{1}{2}g^{i(3)}_j R - (K^i_j - \delta^i_j K)_{,\eta} - KK^i_j + \frac{1}{2}\delta^i_j[K^2 + K_{ij}K^{ij}] \\ = \kappa T^i_j, \end{aligned} \tag{110}$$

where $\kappa := 8\pi G$, and a slash denotes the covariant derivative with respect to the induced hypersurface metric $h_{\mu\nu}$.

Besides the Einstein equations (108) - (110), there are further restrictions on the bubble-wall stress-energy tensor which follow from the covariant stress-energy conservation equations. These equations take the form [1],

$$T^{i\nu}_{;\nu} = T^{i\eta}_{,\eta} + 2K^i_j T^{j\eta} + KT^{i\eta} = 0, \tag{111}$$

$$T^{\eta\nu}_{;\nu} + T^{\eta i}_{;i} + T^{\eta\eta}_{,\eta} - K_{ij}T^{ij} + KT^{\eta\eta} = 0. \tag{112}$$

Substituting the general expression for the stress-energy tensor (107) into the conservation equation (111) yields,

$$T^{i\nu}_{;\nu} = [S^{ij}_{|j} + 2K^i_j S^{j\eta} + KS^{i\eta}]\ \delta(\eta) + S^{i\eta}\delta'(\eta) = 0, \tag{113}$$

where a prime denotes differentiation with respect to η. By requiring that the factors which multiply $\delta(\eta)$ and $\delta'(\eta)$ vanish separately, we obtain the conditions

$$S^{i\eta} = 0, \tag{114}$$

and

$$S^{ij}_{;j} = 0. \tag{115}$$

Following the same argument, the conservation equation (112) yields the condition,

$$T^{\eta\nu}_{;\nu} = [-\Lambda - \frac{1}{2}(K_{ij}(0+) + K_{ij}(0-))S^{ij} + KS^{\eta\eta}]\delta(\eta)\ + S^{\eta\eta}\delta'(\eta) = 0, \tag{116}$$

where $K_{ij}(0+) := \lim_{\eta\downarrow 0} K_{ij}(\eta)$ and $K_{ij}(0-) := \lim_{\eta\uparrow 0} K_{ij}(\eta)$. Since the factors which multiply $\delta(\eta)$ and $\delta'(\eta)$ must vanish separately in equation (116), it follows that

$$S^{\eta\eta} = 0. \tag{117}$$

An equation of motion for the bubble-wall follows by integrating the Einstein equation (110) over an infinitesimal η interval about $\eta = 0$. When we perform this integration, it is important to note that K^i_j and $R^{(3)}$ are everywhere finite, and hence the integral of linear combinations of these quantities over an infinitesimal η-interval vanishes. Hence, the only term which contributes to the integral over η on the left-hand side of equation (110) is the term which is a total derivative with respect to η. This term is integrated trivially, and yields the difference of K^i_j at both sides of the bubble-wall. Similarly, by integrating the right-hand side of equation (110) over an infinitesimal η-interval, we extract only a term which multiplies a δ-function in the expression for the stress-energy tensor (107). Hence, we obtain,

$$K^i_j(0-) - K^i_j(0+)\ = \kappa(S^i_j - \frac{1}{2}\delta^i_j S). \cdot \tag{118}$$

The extrinsic curvatures of the bubble-wall which appear in expression (118) can be calculated from the expression for the metric at both sides of the bubble-wall, which yields a relation between the bubble-wall radius r and first and second-order time derivatives thereof, and the bubble-wall stress-energy S^i_j. We will not repeat this calculation here, which can be found, *e.g.*, in [1, 13]. Instead, we will focus on the *inconsistency* which occurs when one tries to reconcile the restrictions which apply to the bubble-wall stress-energy tensor with the stress-energy tensor of a realistic thin bubble-wall.

Let us now recall the explicit expression for the scalar field stress-energy tensor, namely,

$$T_{\mu\nu} = \partial_\mu\phi\partial_\nu\phi \ - \frac{1}{2}g_{\mu\nu}(\partial_\mu\phi\partial^\mu\phi \ + V(\phi)). \tag{119}$$

Since the bubble-wall forms the boundary between a false-vacuum region where ϕ is equal to a nonzero constant, say ϕ_0, and a Minkowski region where ϕ vanishes, it follows that $\phi = \phi_0$ for $\eta > 0$ and $\phi = 0$ for $\eta < 0$. The fact that ϕ depends only on the η-coordinate ensures that the components of the gradient of ϕ which are tangent to the bubble-wall must vanish. The component of the gradient of ϕ in the direction of the vector $\partial/\partial\eta$ is easily identified as the distribution $\frac{\partial}{\partial\eta}\phi_0\theta(\eta) = \phi_0\delta(\eta)$.

The vanishing of gradients of the scalar field in the directions tangent to the bubble-wall is indeed confirmed by combining condition (114) with the expression for the stress-energy tensor (77). Note, however, that the potential term in expression (119) for $T_{\mu\nu}$ does not contain a δ-function, such that the δ-function part of $T_{\mu\nu}$ can only be generated by the gradient of the ϕ-field, *i.e.*,

$$S_{\eta\eta} = \frac{1}{2}\partial_\eta\phi\partial_\eta\phi, \tag{120}$$

and

$$S_{ij} = -\frac{1}{2}g_{ij}\partial^\eta\phi\partial_\eta\phi, \tag{121}$$

where we used that $g_{\eta\eta}g^{\eta\eta} = 1$. According to equations (120) and (121) it follows that $S_{\eta\eta}$ can only vanish when S_{ij} vanishes as well. The vanishing of all components of $S_{\mu\nu}$ is inconsistent with the assumed step-function behavior of the scalar field at the bubble-wall, and it appears that a non-vanishing bubble-wall tension S_{ij} is necessary to obtain a non-trivial equation of motion (118) for the bubble-wall, which is our main objection against the thin-wall method.

Another point which needs to be mentioned is that the derivative of the potential $V(\phi)$ enters the equation of motion (83) with a negative sign in the region where the gradient of the scalar field is spacelike. Hence, one expects that the scalar field evolves to *higher* values of the potential, as a function of the comoving time parameter, in the region of the spacetime where the constant-ϕ hypersurfaces are timelike and expanding. Similarly, the decay of the scalar field towards the true minimum of the potential can be expected to occur most significantly (but not exclusively) in the region of the spacetime where the constant-ϕ hypersurfaces

are spacelike. In the thin-wall approach the region where the scalar field decays is idealized by a timelike hypersurface, and hence this method cannot be expected to be accurate (if it is relevant at all) to describe the realistic decay of the scalar field in the region where the constant-ϕ hypersurfaces are spacelike.

Another problematic point of the thin-wall approach is that one has to *assume* a specific geometry which forms the interior of the bubble (*e.g.*, Minkowski or Schwarzschild spacetime). Hence, we have constructed a spacetime in which the stress-energy of the bubble-wall cannot dissipate into the interior of the bubble, since here the stress-energy tensor is required to vanish. It is therefore not surprising that we have found that the bubble-wall stress-energy is covariantly conserved *within* the bubble-wall, which is expressed by equation (115). Since the bubble-wall stress-energy enters the equation of motion for the bubble-wall through equation (118), the dynamics of the bubble-wall will also depend on the *a priori* assumed form of the geometry in the interior of the bubble. Clearly, this dependence on *a priori* assumptions limits the relevance of the thin-wall spacetime as a model for a realistic bubble spacetime.

Finally, one may question the assumption that the induced three-metric g_{ij} at a constant-η hypersurface is continuous at $\eta = 0$. In [1] it is argued that the continuity of g_{ij} implies finiteness of K_{ij}, and hence the field equation (108) and the definition (107) imply that $S^{\eta\eta}$ must vanish. This argument does not seem to be solid, in the sense that K_{ij} may contain terms which are proportional to $\delta^{\frac{1}{2}}(\eta)$. The addition of a term of this type to K_{ij} does not change the continuity of the induced three-metric $h_{ij}(\eta)$ about $\eta = 0$. In order to show this, let us note that by integrating expression (8) combined with the definition (3) in the appendix 4.B, it follows that

$$h_{ii}(\eta) = h_{ii}(\eta_0)\ \exp[2\int_{\eta_0}^{\eta} \mathrm{d}\eta K_i], \tag{122}$$

where $K_i := K_{ii}h^{ii}$ and no summation over i is implied. It is clear that expression (122) remains unchanged by adding a term to K_{ij} which is proportional to $\delta^{\frac{1}{2}}(\eta)$, since the integral over η of $\delta^{\frac{1}{2}}(\eta)$ vanishes. Note, however, that the right-hand side of the field equation (108) picks up a term proportional to $\delta(\eta)$ when we add a term proportional to $\delta^{\frac{1}{2}}(\eta)$ to K_{ij}. Since this ambiguity is related to the fact that the extrinsic curvature enters the constraint equation (89) quadratically, while the energy density enters the equation linearly, this effect seems to be an extreme manifestation of the averaging problem in general relativity (see chapter 2 in this thesis).

4.5 Exact spherically symmetric bubble dynamics

In the previous section we discussed the thin-wall description of bubble-dynamics. Although most investigations of bubble dynamics appear to be based on the thin-wall approach, we have presented several reasons why this method can be questioned. As an alternative, one could consider a generic $O(1,3)$ symmetric bubble-

spacetime, and as we have shown in the beginning of this chapter, the dynamics of this spacetime is described by two coupled second-order differential equations for two scalar variables. Although this set of equations is easy to solve by numerical methods, it appears that the restriction to $O(1,3)$ symmetric solutions excludes all realistic types of bubble geometries. One reason for this is that an $O(1,3)$ symmetric bubble is time-reversal invariant, and hence it cannot describe the situation where a bubble is created at some time. Further, the construction of an $O(1,3)$ symmetric bubble requires that one assumes from the start that the spacetime has the same global structure as the De Sitter hyperboloid $\mathcal{H}$. For these reasons, it seems of interest to formulate the set of equations which describes the dynamics of a generic bubble-spacetime, which, for simplicity, we assume to be either spatially spherically symmetric, or translation invariant in two spatial directions. The set of equations which describes the dynamics of this geometry seems too complicated to be solved analytically, but it seems rather simple to solve these equations by numerical methods. For reasons of finite time we have not performed this numerical calculation. The rest of this section is rather technical, and it is probably not of much interest for most readers, except for those who pursue a numerical implementation of the equations which we derive.

The main difficulty consists of formulating a first-order evolution equation for the trace of the extrinsic curvature, K, in terms of the spatial metric components and ϕ and $\dot{\phi}$. By using the momentum constraint equation, we can determine the traceless part of the extrinsic curvature, from the trace part. Therefore, the first-order evolution equation for K yields a first-order evolution equation for the extrinsic curvature tensor, which is a second-order evolution equation for the spatial metric. Further, we have a second-order evolution equation for the scalar field ϕ, (83). These two coupled equations determine the time evolution of the scalar field and the spatial metric, provided that the initial values for $\phi, \dot{\phi}, K$, and the two independent spatial metric components are specified on an initial hypersurface.

4.5.1 Variables and initial conditions

The line element of a general spatially spherically symmetric spacetime, in terms of a comoving coordinate system, has the form

$$\mathrm{d}s^2 = g_{00}\mathrm{d}t^2 + g_{11}\mathrm{d}r^2 \ + g_{22}\mathrm{d}\Omega^2, \tag{123}$$

where $\mathrm{d}\Omega^2 = \mathrm{d}\theta^2 + \sin^2\phi \mathrm{d}\phi^2 (\phi \in [0, 2\pi], \theta \in [0, \pi])$ denotes the line element on the two-dimensional unit sphere, or, without changing the essence of our calculation, a Euclidean two-plane, in which case $\mathrm{d}\Omega^2 = \mathrm{d}\theta^2 + \mathrm{d}\phi^2$.

It should be noted that the metric component g_{00} is not an independent variable, since the comoving time coordinate x^0 is by definition constant on hypersurfaces on which ϕ is constant. Hence, it follows that $\frac{\partial}{\partial x^0}\phi$ is spatially constant, and since

$$\dot{\phi}^2 = g^{00}(\frac{\partial \phi}{\partial x^0})^2, \tag{124}$$

it follows that

$$g_{00} = \frac{c(t)}{\dot{\phi}^2}, \tag{125}$$

where $c(t)$ is an arbitrary function of time which is related to the re-parameterization freedom which is present in the choice of a comoving time coordinate.

The induced metric h_{ij} on the hypersurfaces of constant comoving time coordinate follows directly from the definition (82) and expression (123), *i.e.*, $h_{ij} = \text{diag.}(g_{11}, g_{22}, g_{22}\sin^2\phi)$. Vice-versa, it follows that the four-dimensional metric $g_{\mu\nu}$ is determined by h_{ij} and $\dot{\phi}$.

The six independent variables which determine the dynamics of our bubble geometry are the two independent components of the spatial metric h_{ij} and the extrinsic curvature K_{ij}, as well as ϕ and $\dot{\phi}$, and these variables depend on the comoving time coordinate x^0, and the radial coordinate x^1. The time evolution of the bubble geometry and the scalar field is determined by two second-order evolution equations for h_{ij} and ϕ, respectively. The evolution equation for ϕ has already been derived in section 4.2, and it is given by equation (83). In the following subsection, we will derive a first-order evolution equation for K, which is a second-order evolution equation for h. The next step is to use the momentum constraint equation to determine K_{ij} in terms of K. The initial data which determines solutions of the geometrical evolution equation consists of h_{ij} and K, as well as ϕ and $\dot{\phi}$, which are specified as a function of the radial coordinate x^1 at a hypersurface of constant time x^0.

In order for the initial data to be consistent with a solution of Einstein's equations, the variables h_{ij}, K_{ij}, ϕ, and $\dot{\phi}$, which are specified on a hypersurface of constant comoving time, must satisfy the Hamiltonian constraint equation. The Hamiltonian constraint equation is given by expression (89) in section 4.2, while the energy density ρ, which appears in this equation is given in terms of ϕ and $\dot{\phi}$ by expression (80). Further, the spatial Ricci scalar can be expressed in terms of the metric components h_{11} and h_{22}. By a rather tedious calculation we obtain,

$$R^{(3)} = \frac{1}{2}h^{11}\frac{(h^{22,1})^2}{(h_{22})^2} - \frac{3}{2}\frac{h^{,1}_{22,1}}{h_{22}} - g^{11}_{,1}\frac{h_{22,1}}{h_{22}}. \tag{126}$$

Given the initial spatial metric h_{ij} for $x^0 = 0$, then, by solving the evolution equation for the extrinsic curvature K_{ij}, we may determine the spatial metric h_{ij} as a function of x^0. Namely, using the definition of the extrinsic curvature (3) in appendix 4.B, it follows that

$$K_i := h^{ii}K_{ii} = \frac{1}{2}\mathcal{L}_n \ln|h_{ii}|. \tag{127}$$

Next, one can use the expression for the Lie-derivative of a scalar (8) in order to integrate expression (127), which yields,

$$\frac{h_{ii}(t)}{h_{ii}(0)} = \exp[2\int_0^{x^0=t_0} \mathrm{d}x^0 N K_i], \tag{128}$$

where $N := (g_{00})^{\frac{1}{2}}$, $K_i := K_{ii}h^{ii}$ where no summation over i is implied. In the following subsection we derive the evolution equation for the extrinsic curvature.

4.5.2 Time evolution of K_{ij}

A formula which describes the time evolution of the trace of the extrinsic curvature, K, has been derived by Ehlers in [14],

$$\dot{K} = R_{\mu\nu}n^\mu n^\nu + \dot{n}^\mu_{;\mu} - \frac{1}{3}K^2 - 2\sigma_{\mu\nu}\sigma^{\mu\nu}, \tag{129}$$

where $\sigma_{\mu\nu} := K_{\mu\nu} - \frac{1}{3}h_{\mu\nu}K$, and $\dot{K} := K_{;\mu}n^\mu$.

The explicit calculation of the various terms on the right-hand side of equation (129) in terms of the metric variables h_{ij} and K_{ij}, as well as the field variables ϕ and $\dot{\phi}$, appears to be rather technical, and they have been worked out in the appendix 4.A. We will just state the results in the following.

The first term on the right-hand side of equation (129) can be related to the matter content of the spacetime by using Einstein's equation and the expression for the scalar field stress-energy tensor. This yields,

$$R_{\mu\nu}n^\mu n^\nu = \dot{\phi}^2 - \frac{1}{2}V(\phi). \tag{130}$$

The term $\dot{n}^\mu$ in equation (129) can be interpreted as the four-acceleration of the normals to the constant ϕ-hypersurfaces. This term can be related to the metric component g_{00}, which is determined by the field variable $\dot{\phi}$ by equation (125). We obtain,

$$\dot{n}^\mu_{;\mu} = -\frac{1}{2}(\ln|\dot{\phi}|)\big|^{|1}_{|1}, \tag{131}$$

where a slash denotes the covariant derivative with respect to the induced metric h_{ij}. The traceless part of the extrinsic curvature, σ_{ij}, which enters the right-hand side of equation (129), can be related to the trace of the extrinsic curvature, K, by using the momentum constraint equation. The resulting equation has the form of a first-order differential equation in K and a variable ζ, *i.e.*,

$$K_{|1} = \zeta_{|1} - \frac{3}{2}\zeta(\ln\frac{h_{11}}{h_{22}})_{|1}, \tag{132}$$

where ζ determines the traceless part of the extrinsic curvature through the relation $\sigma^i_j = \zeta\mathrm{diag}(\frac{2}{3}, -\frac{1}{3}, -\frac{1}{3})$. The unappealing aspect of equation (132) is that in order to determine σ_{ij} as a function of x^1, in terms of K and h_{ij}, one needs to integrate the first-order differential equation from the origin of the spherical coordinate system, where $\zeta = 0$, to arbitrary values of the radial coordinate x^1.

4.5.3 Discussion

We derived the two coupled second order differential equations which describe the dynamics of a general spherically symmetric 3 + 1 dimensional geometry which is coupled to a scalar field with a potential term. These equations are well defined as long as the constant-ϕ hypersurfaces are not null. The complexity of these equations suggests that a numerical approach is needed to solve them.

4.6 Open inflation without false-vacuum

In the previous two sections we discussed the possibility that an open FLRW geometry is generated by the decay of a false-vacuum state. In this section we will investigate the possibility that a negatively curved spacetime is generated through inflation without assuming the existence of a false-vacuum state. Further, we do not constrain the initial conditions for the scalar field and the geometry in the sense that we assume that the scalar field can take arbitrary values on an arbitrary initial three-surface. This type of initial conditions are called 'chaotic', and they where first discussed in the context of inflation in [15]. The precise method by which we specify arbitrary initial data is not important in the discussion which follows, but, since our approach is based on classical general relativity and classical field theory, our description is limited to the regions of the spacetime where these theories are accurate. We should note that the aim of this section is to present some more speculative ideas concerning the structure of an inflating spacetime which is subject to chaotic initial conditions. Rather unexpectedly, we find that the initial data which determines the scalar field and the geometry at late times are naturally constrained, in spite of the random nature of the initial conditions. More specifically, it is shown that negative spatial curvature is generated naturally when the hypersurfaces on which the scalar field is constant make a transition from being timelike to being spacelike at some time in the past.

Let us consider two subclasses of chaotic initial conditions, depending on whether or not the gradient of the scalar field is almost everywhere timelike. In this definition 'almost everywhere' means at all subsets of our spacetime which have a non-vanishing four-dimensional Lebesgue measure.

In the first case, *i.e.*, when the gradient of the scalar field is almost everywhere timelike, one can introduce a comoving coordinate system, which we introduced in section 4.2, such that the hypersurfaces of constant comoving time coordinate are everywhere spacelike. Since the ϕ-field is by definition constant on hypersurfaces of constant comoving time coordinate, the randomness of the initial data is present only in the form of the random *geometrical* properties of an initial comoving spatial hypersurface. This situation has been discussed in section 4.2, where we found that the spacetime expands nearly exponentially as long as the spatial curvature is small compared to the energy density and the square of the shear of the normals. Under these conditions, the spatial curvature of the spacetime is generated by the initial

inhomogeneity of the hypersurfaces on which the scalar field is constant, and hence this type of inflation cannot be used to explain the homogeneity *and* the negative spatial curvature in the observed universe.

The second class of chaotic initial conditions is defined by the condition that the gradient of the scalar field is not almost everywhere timelike, and we will study this case in the remaining part of this section.

Let us first introduce some definitions. We denote our 3+1 dimensional spacetime by the symbol $\mathcal{M}$. For simplicity, we assume that the scalar field configuration is differentiable, such that its gradient $\partial_\mu \phi$ is well defined. We can write $\mathcal{M}$ as the union of the following subspaces,

$$\begin{aligned} \mathcal{T} &:= \{\mathcal{M} | \partial_\mu \phi \quad \text{is timelike}\}, \\ \mathcal{S} &:= \{\mathcal{M} | \partial_\mu \phi \quad \text{is spacelike}\}, \\ \mathcal{N} &:= \{\mathcal{M} | \partial_\mu \phi \text{ is null or vanishing}\}, \end{aligned} \tag{133}$$

where $\partial_\mu \phi$ is defined to be timelike, spacelike, or null or vanishing when $\partial_\mu \phi \partial^\mu \phi$ is smaller than zero, greater than zero, or equal to zero, respectively. Indeed, it follows trivially from the definition (133) that

$$\mathcal{M} = \mathcal{T} \cup \mathcal{S} \cup \mathcal{N}, \tag{134}$$

when $\partial_\mu \phi$ is well defined at all points in $\mathcal{M}$.

The assumption that the gradient of the scalar field is not almost everywhere timelike implies that $\mathcal{S}$ is a four-dimensional subspace of $\mathcal{M}$. Further, if one requires that $\mathcal{M}$ includes a region that could describe a physically realistic universe, which we define by the condition that the scalar field evolves to a minimum of the potential at late times, then $\mathcal{M}$ must contain a four-dimensional subspace $\mathcal{T}$. Since it is our aim to describe a spacetime which includes a physically realistic universe, we will make this additional assumption. Let us note that it follows from the definition (133) that the two subspaces $\mathcal{T}$ and $\mathcal{S}$ do not intersect, *i.e.*,

$$\mathcal{T} \cap \mathcal{S} = \emptyset, \tag{135}$$

and it follows from the definition (133) that $\mathcal{T}$ and $\mathcal{S}$ are both open subspaces in $\mathcal{M}$. Let us introduce the boundary $\mathcal{B}$ of $\mathcal{T}$, which is defined as the collection of points which are in the closure of $\mathcal{T}$, but which are not in $\mathcal{T}$. Since $\mathcal{T}$ is a four-dimensional subspace of $\mathcal{M}$ it follows that $\mathcal{B}$ is a three-dimensional compact hypersurface in $\mathcal{M}$, or a collection of three-dimensional compact hypersurfaces in $\mathcal{M}$. It follows trivially from this definition that $\mathcal{B}$ does not intersect with $\mathcal{T}$. Note that $\mathcal{B}$ is also the boundary of the complement of $\mathcal{T}$ in $\mathcal{M}$, which is a closed subspace of $\mathcal{M}$. Since $\mathcal{S}$ is an open subspace of $\mathcal{M}$, it follows that $\mathcal{B}$ cannot intersect with $\mathcal{S}$, namely, if $\mathcal{S}$ would intersect with $\mathcal{B}$ then this would imply that $\mathcal{S}$ is a closed subspace of $\mathcal{M}$ at those points where $\mathcal{S}$ intersects with $\mathcal{B}$. Hence, it follows from equation (134) that $\mathcal{B}$ must be contained in $\mathcal{N}$.

We distinguish two subspaces of $\mathcal{B}$. The first subspace of $\mathcal{B}$, which we call $\mathcal{B}^0$, consists of those points $p \in \mathcal{B}$ for which $\partial_\mu\phi$ vanishes, while the hypersurface $\mathcal{B}$ is spacelike at p. It appears that the hypersurfaces $\mathcal{B}^0$ are of no relevance in our derivation, since the induced metric of the constant-ϕ hypersurfaces appears to be regular at $\mathcal{B}^0$. The rest of this paragraph is devoted to deriving this result.

Let τ denote the proper time τ which is measured along integral curves of $\partial_\mu\phi$ in $\mathcal{T}$, and we choose the origin of the τ-coordinate such that $\tau = 0$ corresponds to the point where integral curves of $\partial_\mu\phi$ intersect $\mathcal{B}^0$. It follows directly from the field equation (83) that $\mathcal{B}^0$ consists of points in $\mathcal{M}$ at which the ϕ-field reaches an extrema, when considered as a function of τ. Further, since $\dot{\phi} = 0$ and $\ddot{\phi} = -V_{,\phi}$ at $\mathcal{B}^0$, it follows that $\dot{\phi}$ has the same sign as $-V_{,\phi}$ ($V_{,\phi}$) for $\tau > 0$ ($\tau < 0$). Hence, since $\dot{V} = \dot{\phi}V_{,\phi}$ is negative (positive) for $\tau > 0$ ($\tau < 0$) it follows that $V(\phi)$ reaches a local *maximum* of the potential at $\mathcal{B}^0$ where $\tau = 0$. This situation is expected to occur when the ϕ-field oscillates about a minimum of the potential, and since $\mathcal{B}^0$ is spacelike, it follows that the constant-ϕ hypersurfaces are spacelike both for $\tau > 0$ and $\tau < 0$. Since the integral curves of $\partial_\mu\phi$ are everywhere well defined in $\mathcal{T}$, one finds that these curves can be extended from points in $\mathcal{T}$ to points at $\mathcal{B}^0$, from which we can extend these integral curves into another section of $\mathcal{T}$. It is of interest to note that the geometry of the constant-ϕ hypersurfaces is nonsingular at $\mathcal{B}^0$, in the sense that the comoving spatial volume, $\sqrt{|h|}$, where $h := \det(h_{ij})$, does not pass through zero at these points. This is clear from the fact that the normals to a spacelike hypersurface $\mathcal{B}^0$ must be well defined and timelike. Hence, it follows from the definition (82) that the induced metric at $\mathcal{B}^0$ is well defined and nonsingular, when the four-dimensional metric is well defined and nonsingular. Hence, these hypersurfaces are of no importance in the following discussion, where we consider nontrivial constraints which apply to the scalar field and the geometry at points of $\mathcal{B}$ where the induced metric of the constant-ϕ hypersurfaces degenerates.

The second subspace of $\mathcal{B}$, which we call $\mathcal{B}^*$, consists of points which are in $\mathcal{B}$, but which are not in $\mathcal{B}^0$. It follows trivially from this definition that $\mathcal{B}^*$ cannot be a spacelike hypersurface on which $\partial_\mu\phi$ vanishes. Further, it follows that $\mathcal{B}^*$ cannot be a timelike hypersurface on which $\partial_\mu\phi$ vanishes, since in this case $\mathcal{B}^*$ must be a timelike constant-ϕ hypersurface. A timelike constant-ϕ hypersurface can only be embedded in $\mathcal{S}$, and this type of hypersurface cannot be the boundary of $\mathcal{T}$ (the proof of this statement, which assumes that $V_{,\phi} \neq 0$, goes identical to the proof above where we show that a spacelike hypersurface on which $\partial_\mu\phi$ vanishes can only be embedded in $\mathcal{T}$). Combining the definition of $\mathcal{B}^*$ with this additional restriction, it follows that $\mathcal{B}^*$ must be a three-dimensional unrestricted hypersurface in $\mathcal{M}$ on which $\partial_\mu\phi$ is null, or $\mathcal{B}^*$ must be locally null while $\partial_\mu\phi$ vanishes.

4.6.1 Comoving dynamics in $\mathcal{T}$

In this section we investigate the behavior of integral curves of $\partial_\mu\phi$ in $\mathcal{T}$, which intersect $\mathcal{B}^*$. In order to be completely general, we will give a description of

all distinct possibilities. It is shown that whenever integral curves of $\partial_\mu\phi$ in $\mathcal{T}$ intersect $\mathcal{B}^*$, or end at $\mathcal{B}^*$, then the induced metric at the comoving hypersurfaces degenerates. This result will be used in the following sections, where we derive nontrivial constraints which apply to the initial data for the scalar field and the geometry in $\mathcal{T}$. There are four distinct ways in which integral curves of $\partial_\mu\phi$ in $\mathcal{T}$ can intersect or end at $\mathcal{B}^*$.

Case 1 occurs when integral curves of $\partial_\mu\phi$ intersect $\mathcal{B}^*$. According to the definition of an integral curve, which is given in the appendix 4.B, this can only happen at those points of $\mathcal{B}^*$ where $\partial_\mu\phi$ is not a tangent vector to $\mathcal{B}^*$. It is clear that integral curves of $\partial_\mu\phi$ can only intersect $\mathcal{B}^*$ at points where this condition is satisfied. Since $\partial_\mu\phi$ is a null vector at $\mathcal{B}^*$, this condition is satisfied automatically at points where $\mathcal{B}^*$ is spacelike. Note that in the case where $\mathcal{B}^*$ is locally a null surface, there will be one null vector which is both normal and tangent to $\mathcal{B}^*$, and in the case where $\mathcal{B}^*$ is locally timelike there will be an $\mathbf{S}^1$ family of null vectors which are tangent to $\mathcal{B}^*$ (*i.e.*, the family of null vectors which span a light-cone in a 2+1 dimensional spacetime; see also the discussion in appendix 4.B). Hence, it follows that when $\mathcal{B}^*$ is locally null or timelike, the condition that $\partial_\mu\phi$ is not tangent to $\mathcal{B}^*$ provides a nontrivial restriction. Whenever integral curves of $\partial_\mu\phi$ intersect $\mathcal{B}^*$, then the constant-ϕ hypersurfaces become locally null, and the induced metric h_{ij} at the constant-ϕ hypersurfaces, which is defined by equation (82), degenerates at these points.

Case 2 occurs when integral curves of $\partial_\mu\phi$ are tangent to $\mathcal{B}^*$. This can only happen at points where $\partial_\mu\phi$ is tangent to $\mathcal{B}^*$. In this case it follows that integral curves of $\partial_\mu\phi$ will be contained in $\mathcal{B}^*$. Since integral curves cannot intersect, this observation implies that integral curves of $\partial_\mu\phi$ in region $\mathcal{T}$ cannot intersect $\mathcal{B}^*$ at points where $\partial_\mu\phi$ is tangent to $\mathcal{B}^*$.

Case 3 occurs at points where $\mathcal{B}^*$ is locally a null surface and $\partial_\mu\phi$ vanishes at $\mathcal{B}^*$, such that ϕ is constant at $\mathcal{B}^*$. It is clear that in this case the integral curves of $\partial_\mu\phi$ are not well defined at $\mathcal{B}^*$. We have not established whether integral curves of $\partial_\mu\phi$ can end in points where $\mathcal{B}^*$ is a null surface and $\partial_\mu\phi$ vanishes. The observation that $\mathcal{B}^*$ is a constant-ϕ hypersurface implies that whenever integral curves of $\partial_\mu\phi$ in $\mathcal{T}$ do end at points at $\mathcal{B}^*$ where $\partial_\mu\phi$ vanishes, then the induced metric h_{ij} degenerates synchronously with respect to the comoving time coordinate.

Case 4 occurs at those points of $\mathcal{B}^*$ which are singular, in the sense that the normals to $\mathcal{B}^*$ in $\mathcal{M}$ are not well defined (*e.g.*, a conical singularity). One expects that a subspace of $\mathcal{B}^*$ which consists of singular points has a dimensionality which is less than three. Let us now consider a family of integral curves of $\partial_\mu\phi$ in $\mathcal{T}$ which intersect or end at the singular subspace of $\mathcal{B}^*$. We distinguish two possibilities, depending on whether or not the family of integral curves which end at singular points at $\mathcal{B}^*$ has a nonzero measure (the measure of a family of curves is given by the Lebesgue measure of the intersection of these curves with a constant-ϕ hypersurface in $\mathcal{T}$). In the case where the family of integral curves has a nonzero

measure, it follows that a spatial volume element of a constant-ϕ hypersurface in $\mathcal{T}$ contracts by an infinite amount as it evolves along integral curves of $\partial_\mu\phi$ which end at singular points at $\mathcal{B}^*$. This situation occurs in the $O(1,3)$ symmetric bubble spacetime which we discussed in the previous section. The second case is not of interest, since in this case only a zero measure subspace of a constant-ϕ hypersurface in $\mathcal{T}$ evolves from singular points at $\mathcal{B}^*$.

By combining these results, it follows that whenever integral curves of $\partial_\mu\phi$ in $\mathcal{T}$ can be extended to points of $\mathcal{B}^*$, then the induced metric h_{ij} at the constant-ϕ hypersurfaces degenerates at these points.

In the following we will focus our interest on the case where there exists a four-dimensional subspace $\mathcal{S}$ in $\mathcal{M}$, and $\mathcal{B}^*$ forms the boundary between two regions $\mathcal{T}$ and $\mathcal{S}$. We should note that $\mathcal{T}$ may contain integral curves of $\partial_\mu\phi$ which can be extended to infinite values of the proper length τ, in both directions. This case has already been analyzed in section 4.2, where we found that, under fairly general assumptions, the geometry evolves as in the standard spatially flat inflation scenario. In the case where integral curves of $\partial_\mu\phi$ can be extended from region $\mathcal{T}$ through $\mathcal{B}^*$ into region $\mathcal{S}$, it follows that the induced metric h_{ij} of the constant-ϕ hypersurfaces must change its signature at $\mathcal{B}^*$. Let us now consider the induced metric h_{ij} as a function of the proper length τ which is measured along the integral curves of $\partial_\mu\phi$, and once again we let $\tau = 0$ correspond to the points where these integral curves intersect $\mathcal{B}^*$. A change of the signature of h_{ij} at $\mathcal{B}^*$ implies that $h := \det(h_{ij})$ changes sign at $\tau = 0$. Hence, it follows that h must either pass through zero at $\tau = 0$, or h changes discontinuously from positive to negative values at $\tau = 0$. A discontinuity of h, which implies a discontinuity of the hypersurface metric h_{ij} at $\tau = 0$, might seem unphysical since in this case the extrinsic curvature, which enters the Hamiltonian constraint equation (89), is not well defined. One should note, however, that the Hamiltonian constraint equation is derived from Einstein's equations by assuming a-priori that the extrinsic curvature is well defined. Hence, we cannot use this equation in order to proof that the induced metric at the constant-ϕ hypersurfaces must be continuous as a function of τ. In the following we will consider the situation where h passes through zero continuously at points where integral curves of $\partial_\mu\phi$ cross $\mathcal{B}^*$. Recall that in the case where a family of integral curves of $\partial_\mu\phi$ with a non-vanishing measure in $\mathcal{T}$ can be continued to singular points at $\mathcal{B}^*$, then it follows that $\sqrt{|h|} \downarrow 0$ at these singular points. Hence, it appears to be a generic result that a comoving volume element $\sqrt{|h|}$ goes through zero along integral curves of $\partial_\mu\phi$ in $\mathcal{T}$ which can be continued into $\mathcal{B}^*$. We should stress that this singular behavior of the induced metric h_{ij} is a coordinate singularity, in the sense that one could choose another coordinate system in terms of which the metric is everywhere regular at $\mathcal{B}^*$. Although the singular behavior of the induced metric h_{ij} does not imply that the four-dimensional geometry at $\mathcal{B}^*$ is singular in any sense, there are nontrivial implications for the dynamics of the scalar field as well as the spatial curvature of the constant-ϕ hypersurfaces when

the comoving volume $\sqrt{|h|}$ goes through zero continuously along an integral curve of $\partial_\mu \phi$ which intersects $\mathcal{B}^*$. We will investigate these implications in the following section.

4.6.2 Initial conditions for the geometry in $\mathcal{T}$

In this subsection we derive constraints which apply to the geometry of the constant-ϕ hypersurfaces in $\mathcal{T}$, in the case where the comoving volume $\sqrt{|h|}$ passes through zero continuously on $\mathcal{B}^*$.

Let us introduce a set of three contravariant vectors $\eta_{(i)}$ with components given by h_{ij} in terms of the basis $\mathrm{d}x^j$. The norm of each of the vectors $\eta_{(i)}$ is given by $|\eta_{(i)}| = \sqrt{|h_{ii}|}|\mathrm{d}x^i|$, where no summation over i is implied. Since the coordinates x^i are transported along integral curves of $\partial_\mu \phi$, we can also interpret $|\eta_{(i)}|$ as the length of a geodesic which is contained in a constant-ϕ hypersurface and which connects two integral curves of $\partial_\mu \phi$ which differ by an infinitesimal coordinate interval $\mathrm{d}x^i$. Let $\tilde{\eta}_{(j)}$ be defined as the unit vector which is associated with $\eta_{(j)}$, such that $\eta_{(j)} = |\eta_{(j)}|\tilde{\eta}_{(j)}$. The physical volume which is associated with a comoving coordinate volume is given by

$$\sqrt{|h|}\mathrm{d}^3x = \sqrt{|\det(\eta_{(1)}, \eta_{(2)}, \eta_{(3)})|}\ \mathrm{d}^3x$$

$$= |h_{11}h_{22}h_{33}|^{\frac{1}{2}}\ \sqrt{|\det(\tilde{\eta}_{(1)}, \tilde{\eta}_{(2)}, \tilde{\eta}_{(3)})|}\ \mathrm{d}^3x, \tag{136}$$

where $\mathrm{d}^3x := \mathrm{d}x^1\mathrm{d}x^2\mathrm{d}x^3$. We assume that the spatial hypersurface coordinates x^i can be chosen to be non-degenerate in some interval $\tau \in (0, \delta)$, where $\delta > 0$. More formally, we require that the limit

$$\lim_{\tau \downarrow 0}\ |\det(\tilde{\eta}_{(1)}, \tilde{\eta}_{(2)}, \tilde{\eta}_{(3)})| \tag{137}$$

exists, and is larger than zero. Condition (137) expresses that the tangent vectors $\eta_{(i)}$ do not degenerate as $\tau \downarrow 0$, although the norm of each of these vectors may vanish in this limit. One should note that the freedom which is involved in choosing the three hypersurface coordinates is sufficient to set the three non-diagonal components of h_{ij} equal to zero, in which case it follows that $\det(\tilde{\eta}_{(1)}, \tilde{\eta}_{(2)}, \tilde{\eta}_{(3)})$ is equal to one. Note, however, that the comoving coordinates x^i are constant along integral curves of $\partial_\mu \phi$, and it is therefore clear that the hypersurface coordinates can only be chosen freely at a single constant-ϕ hypersurface. Hence, we cannot in general require that the induced metric h_{ij} is diagonal for more than one value of the comoving time parameter. In the following discussion, we will assume that the hypersurface coordinates x^i are chosen such that the limit (137) is equal to one, which implies that the tangent vectors $\tilde{\eta}_{(i)}$ become orthonormal as $\tau \downarrow 0$. We then investigate the situation where $h \downarrow 0$ along an integral curve of $\partial_\mu \phi$ at some finite value of the proper length $\tau := 0$. Provided that condition (137) holds, it

then follows from expression (136) that h can only vanish when the norm of at least one of the tangent vectors $\eta_{(i)}$ vanishes as $\tau \downarrow 0$. The case where the norm of one of the tangent vectors $\eta_{(i)}$ vanishes as $\tau \downarrow 0$ occurs at points on $\mathcal{B}^*$ where the constant-ϕ hypersurfaces become locally a null-surface, in which case one of the tangent vectors $\eta_{(i)}$ becomes a null vector. As we pointed out in subsection 4.6.1, this situation occurs whenever integral curves of $\partial_\mu \phi$ intersect $\mathcal{B}^*$, and hence it occurs in case 1 and 3 which are discussed in subsection 4.6.1. Another possibility which may occur is that the norm of n of the three tangent vectors $\eta_{(i)}$ vanishes at points where a family of integral curves of $\partial_\mu \phi$ with nonzero measure in $\mathcal{T}$ ends at a $3-n$ dimensional subspace of $\mathcal{B}^*$, where $n \in \{1,2,3\}$ (*i.e.*, case 4 which is discussed in subsection 4.6.1).

Let us now consider the constraints which apply to the geometry of the constant-ϕ hypersurfaces at $\mathcal{B}^*$. It is useful to introduce the variable K_i, which is defined by,

$$K_i := h^{ii} K_{ii} = \frac{1}{2} \mathcal{L}_n \ln |h_{ii}|, \tag{138}$$

where no summation over i is implied. It is clear from expression (138) that K_i diverges whenever $|h_{ii}|$, which equals the squared norm of $\eta_{(i)}$, goes to zero. It follows from the constraint equation (87) that the spatial curvature of the constant-ϕ hypersurfaces in $\mathcal{T}$ satisfies the asymptotic relation

$$R^{(3)} \approx -K^2 + K_{ij} K^{ij}, \tag{139}$$

in the limit where $\tau \downarrow 0$, and we assumed that the energy density ρ remains finite on $\mathcal{B}^*$. Let us first consider the asymptotic behavior of $R^{(3)}$ in the case where only one of the diagonal components of the hypersurface metric, which we label h_{11}, goes to zero in the limit where $\tau \downarrow 0$. Using the definition (138) it follows that

$$K^2 \approx K_1^2 + 2K_1(K - K_1), \tag{140}$$

and

$$K_{ij} K^{ij} \approx K_1^2, \tag{141}$$

to divergent order as $\tau \downarrow 0$. By substituting expressions (140) and (141) in expression (139), we obtain the asymptotic expression,

$$R^{(3)} \approx -2K_1(K_2 + K_3), \tag{142}$$

in the limit where $\tau \downarrow 0$. Interestingly, the spatial curvature of the constant-ϕ hypersurfaces diverges on $\mathcal{B}^*$ and is *negative* when the sum of the expansions in the directions of $\eta_{(2)}$ and $\eta_{(3)}$ is positive. In the case where the norm of more than one of the tangent vectors $\eta_{(i)}$ approaches zero along integral curves of $\partial_\mu \phi$ as $\tau \downarrow 0$, it follows that $R^{(3)}$ diverges as,

$$R^{(3)} \approx -\sum_{i \neq j} K_i K_j, \tag{143}$$

and it is clear that the divergent part of $R^{(3)}$ is negative definite in this case. Hence, it appears that the curvature of the constant-ϕ hypersurfaces is negative and divergent along integral curves of $\partial_\mu\phi$ in $\mathcal{T}$ which can be extended into $\mathcal{B}^*$. In deriving this result, we have assumed that h passes through zero continuously as a function of the proper length along integral curves of $\partial_\mu\phi$ which intersect $\mathcal{B}^*$. In the following subsection we will consider the constraints which apply to the scalar field when $h \downarrow 0$ along integral curves of $\partial_\mu\phi$ in $\mathcal{T}$ which intersect $\mathcal{B}^*$.

4.6.3 Initial conditions for the scalar field in $\mathcal{T}$

In this subsection we show that $\dot{\phi}$ must vanish on $\mathcal{B}^*$, for physically acceptable solutions of the field equation, which satisfy the condition that the energy density ρ is finite on $\mathcal{B}^*$.

Our derivation of this result is based on the energy conservation equation (84), where we substitute expression (91) for the volume expansion K. We obtain,

$$\dot{\rho} = -(\rho + p)(\ln \sqrt{|h|})\dot{}. \tag{144}$$

In our analysis of solutions of the conservation equation (84), we will use that $\rho+p$ is continuous in an interval $\tau = (0, \delta)$, where $\delta > 0$, and once again we choose $\tau = 0$ to be the point for which the comoving volume element $\sqrt{|h|}$ vanishes. More precisely, we will use that there exists some $\delta \in \mathbf{R}^+$ such that $\rho+p$ is continuous for $|\tau| \in (0, \delta)$. In order to obtain this result, let us note that according to expressions (80) and (81), it follows that $\rho + p$ equals $\dot{\phi}^2$. Hence, the field equation (1) shows that $\ddot{\phi}$ must be finite when $V_{,\phi}$ and K are finite for $\tau \neq 0$, which we assume to hold. Finiteness of $\ddot{\phi}$ implies continuity of $\dot{\phi}$, which implies continuity of $\rho+p = \dot{\phi}^2$ with respect to τ derivation for $\tau \neq 0$.

We distinguish two types of solutions of equation (84), depending on whether $\rho + p$, or equivalently $\dot{\phi}^2$, goes to zero in the limit where $\tau \downarrow 0$. Recall that the limit of $\rho+p$ for $\tau \downarrow 0$ exists and is equal to some number c, if and only if for every $\epsilon \in \mathbf{R}^+$ there is a $\delta \in \mathbf{R}^+$ such that

$$|\rho + p - c| < \epsilon, \tag{145}$$

for all $\tau \in (0, \delta)$. When $\dot{\phi}$ is continuous about $\tau = 0$ but $\dot{\phi}$ does *not* approach to zero in the limit where $\tau \downarrow 0$, then it follows from expression (145) that there exists an $\alpha \in \mathbf{R}^+$ such that

$$|\rho + p| > \alpha, \tag{146}$$

for all $\tau \in (0, \delta)$ (we can proof this statement by using that $|\rho+p|-\epsilon < c < |\rho+p|+\epsilon$, and then we choose $\delta, \epsilon, \alpha \in \mathbf{R}^+$ such that condition (145) holds while $\alpha+\epsilon < c$). By using condition (146) in the conservation equation (144), we obtain the inequality,

$$|\dot{\rho}| > |\alpha \ln \sqrt{|h|}|, \tag{147}$$

which holds for all $\tau \in (0, \delta)$. By integrating equation (147) with respect to τ over an interval $(\sigma, \delta]$, where $\sigma \in [0, \delta]$, we obtain

$$|\rho(\delta) - \rho(\sigma)| > \alpha |\ln \sqrt{|h(\delta)|}| - \alpha |\ln \sqrt{|h(\sigma)|}|. \tag{148}$$

By taking the limit $\sigma \downarrow 0$ in equation (148), we find that $\rho(\delta)$ diverges at least as fast as $\ln |h(\sigma)|$ when $\sigma \downarrow 0$, but this is a weak lower bound on the strength of the divergence of ρ, since, as we have shown in appendix 4.C, $\dot{\phi}$ diverges as fast as $|h(\sigma)|^{-1}$ for singular solutions of the field equation in the limit where $\sigma \downarrow 0$. Therefore, finiteness of the energy density ρ implies that $\dot{\phi}$ must approach to zero in the limit where $\tau \downarrow 0$. Indeed, this result appears rather natural when we recall that an integral curve of $\partial_\mu \phi$ in $\mathcal{T}$ becomes null for finite values of the proper length τ. Hence, one expects that the proper acceleration diverges along an integral curve of $\partial_\mu \phi$ which is continued from $\mathcal{T}$ into $\mathcal{B}^*$. Indeed, expression (7) in appendix 4.A shows that a divergence of the proper acceleration implies that $|\dot{\phi}| \downarrow 0$ along these integral curves.

An interesting implication of this result is that in the case where integral curves of $\partial_\mu \phi$ intersect $\mathcal{B}^*$ towards the *future*, then either the energy density on $\mathcal{B}^*$ diverges, or the initial data for the scalar field and the geometry in $\mathcal{T}$ must be such that $\dot{\phi}$ vanishes exactly at $\mathcal{B}^*$. The occurrence of infinite energy densities on $\mathcal{B}^*$, or the need for an infinite amount of fine-tuning of the initial data in $\mathcal{T}$, does not seem to be physically acceptable. Hence, given the assumption of continuity of h on $\mathcal{B}^*$, we find that a physically acceptable solution of the field equation in $\mathcal{T}$ is only compatible with a choice of an arrow of time for which all integral curves of $\partial_\mu \phi$ in $\mathcal{T}$ intersect $\mathcal{B}^*$ towards the past.

4.6.4 Discussion

We considered the dynamics of a scalar field ϕ with a potential term in a 3+1 dimensional geometry, and we assumed random initial conditions for the scalar field and the geometry. A spacetime with this type of initial conditions naturally contains regions where the constant-ϕ hypersurfaces are spacelike and timelike, respectively. The regions of the spacetime where the constant-ϕ hypersurfaces are spacelike have the potential to evolve into a physically realistic universe at late times, and the geometry of the spacetime at late times is determined by the geometry of the constant-ϕ hypersurfaces. It is therefore of interest to determine the initial conditions which apply to the geometry and the scalar field at the hypersurface which bounds the region of the spacetime where the constant scalar field hypersurfaces are spacelike. It is shown that the constant scalar field hypersurfaces start their spacelike spacelike evolution with a singular negative spatial curvature, while the stress-energy is restricted by the condition that $\rho + p$ must vanish at these points. This result is however subject to the assumption that the comoving volume element, $\sqrt{|h|}$, passes through zero continuously on $\mathcal{B}^*$, while we have not

been able to outrule a discontinuous change. In the case where integral curves of $\partial_\mu \phi$ can be continued through $\mathcal{B}^*$ (*i.e.*, case 1 in section 4.6.1), this assumption seems to be rather nontrivial. It should be noted that we have not been able to establish or outrule whether the dynamics of the scalar field allows for the situation where integral curves of $\partial_\mu \phi$ can be continued through $\mathcal{B}^*$. If the answer on this question is positive, and it appears that the induced metric h_{ij} changes its signature discontinuously at $\mathcal{B}^*$, then this could invalidate the result which are derived in the previous two subsections. In order to clarify these points, it is necessary to gain a better understanding of the dynamics of the constant-ϕ hypersurfaces in the region of the spacetime where these hypersurfaces are timelike.

An interesting but highly speculative idea is that a discontinuous signature change of the hypersurface metric could explain the generation of a negatively curved universe which is nearly homogeneous as well as isotropic. Let us therefore observe that it seems rather plausible that the geometry of a timelike constant-ϕ hypersurface in a nearly exponentially expanding spacetime corresponds to the geometry of a 2+1 dimensional spacetime which expands nearly exponentially. If this assumption holds, then the line element is approximately given by

$$\mathrm{d}s^2 \approx -\mathrm{d}\sigma^2 + \exp[2H\sigma]\mathrm{d}E^2, \tag{149}$$

where $\mathrm{d}E^2$ denotes the line element of a two-dimensional Euclidean plane. If a discontinuous change of the signature of the metric of the constant-ϕ hypersurfaces takes place where these hypersurfaces intersect $\mathcal{B}^*$, then it would not be surprising when the line-element of these hypersurfaces takes the form,

$$\mathrm{d}s^2 \approx \mathrm{d}\sigma^2 + \exp[2H\sigma]\mathrm{d}E^2. \tag{150}$$

The line element at the right hand side of equation (150) describes a negatively curved homogeneous and isotropic three-space. Hence, it does not seem to be ruled out that a negatively curved nearly FLRW universe can be generated through exponential expansion, which is followed by a discontinuous change of the signature of the induced metric h_{ij} of the constant-ϕ hypersurfaces.

4.A K-evolution equation

The time evolution of the expansion K is described by the formula [14],

$$\dot{K} = R_{\mu\nu}n^\mu n^\nu + \dot{n}^\mu_{;\mu} - \frac{1}{3}K^2 - 2\sigma_{\mu\nu}\sigma^{\mu\nu}, \tag{1}$$

where $\sigma_{\mu\nu} := K_{\mu\nu} - \frac{1}{3}Kh_{\mu\nu}$, and $\dot{K} := K_{;\mu}n^\mu$. The first term on the right-hand side of equation (1) is related to the matter content of the spacetime by Einstein's equations. By substituting the expression for the scalar field stress-energy tensor (77) into Einstein's equations, *i.e.*,

$$R^\mu_\nu - \frac{1}{2}\delta^\mu_\nu R = \kappa T^\mu_\nu, \tag{2}$$

we obtain an expression for the Ricci tensor in terms of the scalar field,

$$R_{\mu\nu} = T_{\mu\nu} - \frac{1}{2}g_{\mu\nu}T = \partial_\mu\phi\partial_\nu\phi + \frac{1}{2}g_{\mu\nu}V(\phi). \tag{3}$$

By contracting equation (3) with the normal vectors n^μ we obtain directly the desired expression,

$$R_{\mu\nu}n^\mu n^\nu = \dot{\phi}^2 + \frac{\epsilon}{2}V(\phi), \tag{4}$$

where $\epsilon = -1(1)$ in the case where the constant-ϕ hypersurfaces are spacelike (timelike). The term $\dot{n}^a$ in equation (1) can be interpreted as the four-acceleration of the normals to the constant ϕ-hypersurfaces, and this term can be related to the metric component g_{00},

$$\dot{n}^\mu := n^\sigma n^\mu_{;\sigma} = n^\sigma n^\mu_{,\sigma} + n^\sigma n^\rho \Gamma^\mu_{\sigma\rho}$$

$$= \delta^\mu_0 (n^0 n^0_{,0} + n^0 n^0 \Gamma^0_{00}) + \delta^\mu_i n^0 n^0 \Gamma^i_{00} = \delta^\mu_1 \Gamma^1_{00} = \epsilon\delta^\mu_1 (\ln|g_{00}|)_{,1}, \tag{5}$$

where a comma denotes the standard derivative with respect to the coordinate x^μ, and we used that $(n^0)^2 = \epsilon g^{00}$, and $\Gamma^i_{00} = \frac{1}{2}g^{ii}g_{00|i}$. The spatial derivative of the metric component g_{00} which appears in equation (5) is determined in terms of the proper time derivative of the field $\dot{\phi}$. Namely, since the comoving time x^0 is defined to be constant on constant-ϕ hypersurfaces, it follows trivially that $\frac{\partial}{\partial x^0}\phi$ is constant on hypersurfaces of constant comoving time x^0. Since $\dot{\phi}^2 = -g^{00}(\frac{\partial}{\partial x^0}\phi)^2$ it follows that

$$g_{00} = \frac{c(x^0)}{\dot{\phi}^2}, \tag{6}$$

where $c(x^0)$ is an arbitrary time dependent function which reflects the re-parameterization freedom which is present in the choice of comoving time variable. Expression (6) can be used to eliminate the metric component g_{00} in expression (5), and we obtain an expression for the four-acceleration of the normals n^μ in terms of a spatial derivative of $\dot{\phi}$,

$$\dot{n}^\mu = \epsilon\delta^\mu_1 (\ln|g_{00}|)_{,1} = -2\epsilon\delta^\mu_1 (\ln|\dot{\phi}|)_{,1}. \tag{7}$$

By taking the divergence of expression (8), it follows directly that

$$\dot{n}^{\mu}_{;\mu} = 2\epsilon(\ln|\dot{\phi}|)^{|1}_{|1}\,, \tag{8}$$

where a slash is defined as the covariant derivative with respect to the induced hypersurface metric. In the rest of this appendix we will determine the traceless part of the extrinsic curvature in terms of the trace part of the extrinsic curvature, by integrating the momentum constraint equation. The momentum constraint equation provides a relation between spatial derivatives of the extrinsic curvature, *i.e.*,

$$K^{ij}_{\ |j} - h^{ij}K_{|j} = J^i = 0, \tag{9}$$

where $J^i := GT^i_0$, and $J^i = 0$ follows from expression (119). Let us now define

$$K_2 := K^{22}h_{22} = K^{33}h_{33}, \tag{10}$$

and

$$\zeta^{ij} := K^{ij} - K_2 h^{ij}. \tag{11}$$

Note that ζ^{ab} is only nonzero for $i = j = 1$, which follows trivially from expression (11) when we use that the metric is diagonal. We may rewrite the constraint equation in the form

$$\zeta^{ij}_{\ |j} - (2K_2 + \zeta)^{|i} = 0, \tag{12}$$

where $\zeta := \zeta^{ij}h_{ij} = \zeta^{11}h_{11}$. We define the covariant and contravariant unit vector in the x_1 direction,

$$k_i = \delta^1_i(|h_{11}|)^{\frac{1}{2}}\,, \quad k^i = \delta^i_1(|h_{11}|)^{-\frac{1}{2}}\mathrm{sgn}(h_{11}), \tag{13}$$

where we used that $k^i k_i = \mathrm{sgn}(h_{11})$. Contracting equation (12) by k_i, we obtain

$$k_i\zeta^{ij}_{\ |j} - (2K_2 + \zeta)^{|i}k_i = 0, \tag{14}$$

which we may write in the form,

$$(k_i\zeta^{ij} - \zeta k^j)_{|j} - 2K_2^{|i}k_i = \ \zeta^{ij}k_{i|j} - \zeta k^j_{\ |j}. \tag{15}$$

The first term in brackets on the left-hand side of equation (15) vanishes, since,

$$k_i\zeta^{ij} = \delta^j_1(|h_{11}|)^{\frac{1}{2}}\zeta^{11} \ = \delta^j_1(|h_{11}|)^{-\frac{1}{2}}|h_{11}|\zeta^{11} \ = k^j\zeta. \tag{16}$$

Further, the first and the second term on the right-hand side of equation (15) can be evaluated using,

$$\zeta^{ij}k_{i|j} = \zeta^{ij}k_{i,j} \ + \zeta^{ij}k_l\Gamma^l_{ij}$$

$$= \zeta^{11}(|h_{11}|)^{\frac{1}{2}}_{,1} \ + \zeta^{11}k_1\frac{1}{2}h^{11}h_{11,1} \ = \zeta^{11}k_1(\ln|h_{11}|)_{,1}, \tag{17}$$

and

$$k^i_{|i} = k^i_{,i} + k^i \Gamma^j_{ij}$$

$$= (|h_{11}|)^{-\frac{1}{2}}_{,1} + (|h^{11}|)^{\frac{1}{2}} \frac{1}{2} h^{11} h_{11,1} + (|h^{11}|)^{\frac{1}{2}} h^{22} h_{22,1} = k^1 (\ln |h_{22}|)_{,1} . \quad (18)$$

Using equation (16) , (17), and (18), the constraint equation (15) can be written in the form,

$$2K_{2,1} = -\zeta \left(\ln \frac{|h_{11}|}{|h_{22}|} \right)_{,1} . \quad (19)$$

By combining expression (19) with the definition (10) and (11), we obtain a relation between ζ and the trace of the extrinsic curvature, *i.e.*,

$$K_{,1} = \zeta_{,1} - \frac{3}{2} \zeta \left(\ln \frac{|h_{11}|}{|h_{22}|} \right)_{,1} , \quad (20)$$

which is the desired equation.

4.B Lie derivation

In this appendix we derive a useful relation between the divergence of the normals to the constant-ϕ hypersurfaces, the trace of the extrinsic curvature, and the Lie-derivative of the determinant of the induced metric on a three-dimensional hypersurface. Since we need to evaluate Lie-derivatives of tensors, let us recall the general formula [5],

$$\mathcal{L}_n T^{\mu_1 \cdots \mu_n}_{\nu_1 \cdots \nu_m} = n^\sigma T^{\mu_1 \cdots \mu_n}_{\nu_1 \cdots \nu_m ; \sigma} - \sum_{i=1}^{m} T^{\mu_1 \cdots \sigma \cdots \mu_n}_{\nu_1 \cdots \nu_m} n^{\mu_i}_{;\sigma}$$

$$+ \sum_{i=1}^{m} T^{\mu_1 \cdot \mu_n}_{\nu_1 \cdot \sigma \cdot \nu_m} n^\sigma_{;\nu_i} . \quad (1)$$

Let us also recall the decomposition of the metric,

$$g_{\mu\nu} = \epsilon n_\mu n_\nu + h_{\mu\nu}, \quad (2)$$

where $h_{\mu\nu} := g_{\mu\nu} - \epsilon n^\mu n^\nu$, and $\epsilon := -1(1)$ in the case where the hypersurface is spacelike (timelike). Taking the Lie-derivative with respect to n^μ on the right-hand side of equation (2) yields

$$\mathcal{L}_n (\epsilon n_\mu n_\nu + h_{\mu\nu}) = \mathcal{L}_n h_{\mu\nu} := 2K_{\mu\nu}. \quad (3)$$

Similarly, taking the Lie-derivative on the left-hand side of equation (2) yields,

$$\mathcal{L}_n g_{\mu\nu} = n^\sigma g_{\mu\nu;\sigma} + g_{\sigma\nu} n^\sigma_{;\mu} + g_{\mu\sigma} n^\sigma_{;\nu} = n_{\nu;\mu} + n_{\mu;\nu}. \quad (4)$$

By combining the results (3) and (4), we obtain

$$n^{\mu}_{;\mu} = g^{\mu\nu}K_{\mu\nu} = K. \tag{5}$$

It is important to note that $n^{\mu}K_{\mu\nu} = 0$, which follows by taking the Lie-derivative on both sides of the equation $n^{\sigma}h_{\sigma\mu} = 0$. Therefore, one finds,

$$K = g^{\mu\nu}K_{\mu\nu} = h^{\mu\nu}K_{\mu\nu}. \tag{6}$$

In the following, we will discuss some standard results which are used in section 4.6.1. Let us recall that a non-vanishing gradient $\partial_{\mu}\phi$ is orthogonal to the three-dimensional hypersurfaces on which ϕ is constant. This might be considered as trivial in the case where $\partial_{\mu}\phi$ is timelike or spacelike, but in the case where $\partial_{\mu}\phi$ is null this may be less obvious. A vector n^{μ} is defined to be orthogonal to the hypersurface Σ if and only if

$$n_{\mu}\eta^{\mu} = 0, \tag{7}$$

for every non-vanishing vector η^{μ} which is tangent to Σ. Note that if one tangent vector is null, or a linear combination of tangent vectors is null, then this null-tangent vector is proportional to the normal vector n^{μ}, and a surface which has this property is called a null-surface. One should note that the vectors which are tangent or orthogonal to a hypersurface in $\mathcal{M}$ do not need to exist, and one expects that this situation occurs when the hypersurface Σ is singular in some sense. Let us now define an integral curve γ_{η} of a vector field η^{μ} as the mapping from $\mathbf{R}$ to $\mathcal{M}$ which satisfies the condition

$$\mathcal{L}_{\eta}f := \partial_{\lambda}f(\gamma_{\eta}(\lambda)) = \eta^{\mu}\partial_{\mu}f, \tag{8}$$

for all differentiable functions f on $\mathcal{M}$. It can be shown that the integral curves of a non-vanishing and continuous vector field η on $\mathcal{M}$ form a uniquely defined family of non-intersecting curves trough every point in $\mathcal{M}$ (see, *e.g.*, [16]). The integral curves of a vector field η^{μ} which are tangent to a hypersurface Σ are entirely contained in Σ (no proof, but it seems obvious). Now let $\gamma_{\eta}(\lambda)$ be an integral curve of a continuous tangent vector field η, which passes through a point p. Then the Lie-derivative of ϕ with respect to η at p is defined by equation (8), *i.e.*,

$$\mathcal{L}_{\eta}\phi \ = \eta^{\mu}\partial_{\mu}\phi, \tag{9}$$

where it follows from the definition (7) that $\mathcal{L}_{\eta}\phi = 0$ when η is tangent to the hypersurfaces which are orthogonal to $\partial_{\mu}\phi$. Hence it follows from the definition (8) that ϕ is constant along all integral curves of η, and therefore ϕ must be constant on hypersurfaces Σ for which $\partial_{\mu}\phi$ is a normal vector.

4.C Slow roll and fast roll

A well known approximate solution to the field equation (83) holds in the case where $K \gg |V_{,\phi}|$, and $\partial_\mu \phi$ is timelike. If these conditions are satisfied, then the field equation, which we obtained in section 4.2, *i.e.*,

$$\ddot{\phi} + K\dot{\phi} + V_{,\phi} = 0, \tag{1}$$

implies that $\dot{\phi}$ must be small and

$$\dot{\phi} \approx -\frac{V_{,\phi}}{K}\,. \tag{2}$$

Expression (2), and the assumption that $K \gg |V_{,\phi}|$, implies that the first term on the left-hand side of equation (1) can be neglected as compared to the second and the third term. This solution describes the friction dominated motion of the field towards lower or higher values of the potential, and hence this approximation is called the 'slow-roll' approximation. Note that, depending on whether K is positive or negative, equation (2) implies that $V_{,\phi}\dot{\phi}$ is negative or positive, and the field evolves to lower or higher values of the potential, respectively. The solution where the field evolves to higher values of the potential and K is negative is simply the time-reverse of the solution where the field evolves to lower values of the potential and K is positive.

Another interesting case occurs when K is negative, and $-K \gg V_{,\phi}$, and we assume for the moment that $\dot{\phi}$ and $V_{,\phi}$ have opposite signs. In this case there is no slow-roll type of solution of equation (1), since the second and the third term on the left-hand side of equation (1) have the same sign, and they cannot add up to zero. The acceleration term $\ddot{\phi}$ must therefore be of the same order as $-K\dot{\phi}$ or $V_{,\phi}$. Since the field accelerates, it follows that $|\dot{\phi}|$ increases in time, and by time that $|\dot{\phi}| \gg |V_{,\phi}/K|$ the field equation (1) is dominated by the first two terms at the right-hand side. In this situation one can approximate equation (1) by taking only the first two terms, *i.e.*,

$$\ddot{\phi} \approx -K\dot{\phi}. \tag{3}$$

Note that when $K < 0$ and K is constant in time, then $\dot{\phi}$ grows as a function of time as $\exp[-K\tau]$, where τ denotes the proper time which is measured along integral curves of $\partial_\mu \phi$. It follows from this observation that the error which is due to neglecting the term $V_{,\phi}$ in equation (1) goes asymptotically to zero. Note also that the consistency of our approximation, which consists of neglecting the term $V_{,\phi}$ in equation (1), does not depend on the sign of $\dot{\phi}V_{,\phi}$.

Equation (3) can also be integrated in the case where K is an arbitrary function of time. Let us therefore recall the definition of K which is given in the previous appendix,

$$K := h^{\mu\nu}K_{\mu\nu} = \frac{1}{2}h^{\mu\nu}\mathcal{L}_n h_{\mu\nu}. \tag{4}$$

and let us note that $h^{\mu\nu}$ in expression (4) can be written in the form,

$$h^{\mu\nu} = \frac{\delta \ln |h|}{\delta h_{\mu\nu}} \tag{5}$$

where $h := \det(h_{\mu\nu})$ (see, *e.g.*, [5]). By substituting expression (5) for $h^{\mu\nu}$ into equation (4), we obtain,

$$K = \frac{1}{2}\frac{\delta \ln |h|}{\delta h_{\mu\nu}} \mathcal{L}_n h_{\mu\nu} = \frac{1}{2}\mathcal{L}_n \ln |h| = (\ln \sqrt{|h|})^{\cdot} . \tag{6}$$

Substituting expression (6) into equation (3) yields

$$(\ln \dot{\phi})^{\cdot} = -(\ln \sqrt{|h|})^{\cdot} . \tag{7}$$

By integrating equation (7), and then exponentiating both terms in the equation, we obtain,

$$\frac{\dot{\phi}(\tau_0)}{\dot{\phi}(\tau)} = \frac{\sqrt{|h(\tau)|}}{\sqrt{|h(\tau_0)|}}, \tag{8}$$

where τ is defined as the proper time which is measured along integral curves of n^{μ}. Equation (8) shows that there is an approximate solution for which $\dot{\phi}$ diverges when the comoving three-dimensional volume $\sqrt{|h(\tau)|}$ goes to zero continously at some time $\tau = \tau_0$.

References

[1] S. K. Blau, E. I. Guendelman and A. Guth, Phys. Rev. **D 35** 1747 (1987).

[2] G. F. Smoot *et al.*, Ap. J. **L1**, 396 (1992).

[3] M. Bucher, A. S. Goldhaber and N. Turok, Phys. Rev. **D 52**, 3314 (19 95).

[4] S. W. Hawking and N. Turok, Phys. Lett. **B 425**, 25 (1998).

[5] R. M. Wald, *General Relativity* (University of Chicago Press, Chicago, 1984).

[6] R. M. Wald, Phys. Rev. **D 28**, 2118 (1983).

[7] J. D. Barrow and G. Götz, Class. Quantum Grav. **6**, 1253 (1989).

[8] A. Joshua *et al.*, Phys. Rev. **D** 30, 265 (1984).

[9] S. W. Hawking and G. F. R. Ellis, *The large scale structure of space-time* (Cambridge University Press, USA, 1973).

[10] W. Israel, Nuovo Cimento, **B 44**, 4349 (1966).

[11] C. G. Callan Jr., and S. Coleman, Phys. Rev. **D 16**, 1762 (1977).

[12] S. Coleman, Phys. Rev. **D 15**, 2929 (1977).

[13] S. Coleman and F. De Luccia, Phys. Rev. **D 21**, 3305 (1980).

[14] J. Ehlers, Gen. Relativ. Grav. **25**, 1225 (1993) (translation of 1961 article).

[15] A. Linde, Phys. Lett. **129 B**, 177 (1983).

[16] E. A. Coddington and N. Levison, *Theory of Ordinary Differential Equations*, (McGraw-Hill, New York, 1955).

List of results

- It is shown that the geometry interior of the horizon of a Schwarzschild black-hole can be connected to a single asymptotically flat region, without violating time-orientability (chapter 1, section 1).
- By identifying points in the Kruskal manifold, we constructed a solution of Einstein's equations describing a circular cosmic string with deficit angle π (chapter 1, section 3).
- Vice-versa, one can construct a sourceless solution of Einstein's equations by identifying points in geometries describing circular cosmic strings (chapter 1, section 3).
- We derive the unique linearized averaging procedure for metrics, which is the infinite power of any linearized averaging operation for which unperturbed FLRW spacetime is a stable fixed point (chapter 2, section 2.2)
- By applying the linearized averaging operation to a perturbed FLRW geometry, we derived an expression for the leading order gauge-invariant correction to the expansion which is due to the non-linearity of Einstein's equations (chapter 2, section 2.4).
- In the observable universe, the correction to the expansion which is due to the backreaction of geometry and matter perturbations is typically small, of the order of 5 parts in 10^5 (chapter 2, section 2.8).
- It is shown that the variation of the action functional of a scalar field in an open FLRW geometry is generally dominated by a surface term which is evaluated at spatial infinity (chapter 3, section 3.5).
- The degrees of freedom which give rise to this surface term have vanishing norm in the Hilbert space of square integrable functions (chapter 3, section 3.6).
- Without quantifying the zero-norm degrees of freedom, the variational description of classical field theory, as well as the path-integral description of quantum field theory, are not well defined in open spacetimes with supercurvature perturbations (chapter 3, section 3.6).
- The thin-wall description of bubble-spacetime dynamics is inconsistent with the assumption that the stress-energy in this spacetime is generated by a scalar field (chapter 4, section 4.4).
- We derived the two coupled second-order differential equations which describe the exact dynamics of a general spherically symmetric bubble-spacetime (chapter 4, section 4.5).

- It is shown that negative spatial curvature is generated naturally at the hypersurface which bounds the region in an inflating spacetime where the constant scalar field hypersurfaces are spacelike (this result is subject to the unproven assumption of continuity of the induced hypersurface metric, see chapter 4, section 4.6.2).
- It is also shown that the kinetic part of the stress-energy tensor must vanish at the hypersurface which bounds the region in an inflating spacetime where the constant scalar field hypersurfaces are spacelike.

Acknowledgement/Dankwoord

- Boven alles wil ik mijn dank uitspreken aan mijn moeder, die mij op alle mogelijke manieren gesteund heeft tijdens mijn verblijf in Zuid Afrika. Verder zou ik mijn grootmoeder willen bedanken voor haar financiele bijdrage aan mijn onderzoek.
- I would like to thank Prof. George F. R. Ellis for providing me with helpful ideas, as well as creating a stimulating enviroment for doing research.
- Furthermore, I would like to thank Henk van Elst for reading the manuscripts which went into this thesis.
- I am also grateful to Mauro Carfora for inviting me to SISSA (Italy), where part of the work for the second chapter in this thesis was done.
- Tenslotte ben ik dank verschuldigd aan Zeger Hendrikse die mij hielp met de presentatie van dit proefschrift.
- The research was partially financed with funds from FRD (South Africa).