

Cosmology

Volker Perlick (perlick@zarm.uni-bremen.de)

Summer Term 2018, University of Bremen

Lectures: Mon 16–18, NW1, N3130

Thu 12–13, NW1, N3310

Tutorials: Thu 13–14, NW1, N3310

General Relativity text-books with detailed sections on cosmology

W. Rindler: “Relativity” Oxford UP (2001)

H. Stephani: “Relativity” 3rd edition, Cambridge UP (2004)

L. Ryder: “Introduction to General Relativity” Cambridge UP (2009)

Monographs

V. Mukhanov: “Physical foundations of cosmology” Cambridge UP (2005)

G. Ellis, R. Maartens, M. MacCallum: “Relativistic Cosmology” Cambridge UP (2012)

Living Reviews (<http://relativity.livingreviews.org>)

N. Jackson: The Hubble Constant, lrr-2015-2

S. Carroll: The Cosmological Constant, lrr-2001-1

A. Jones and A. Lasenby: The Cosmic Microwave Background, lrr-1998-11

Contents

1	Historic Introduction	2
2	Brief review of general relativity	4
3	Homogeneous and isotropic cosmology	8
3.1	Robertson-Walker spacetimes	8
3.2	Friedmann-Lemaître solutions	29
4	Observations	72
4.1	Evidence for dark matter	72
4.2	The distance-redshift relation	77
4.3	The cosmic background radiation	79
4.4	Other observations	90
5	Perturbation theory	92
6	Bianchi models	108
7	Singularity theorems	119

1 Historic Introduction

- 1826** W. Olbers formulates the “Olbers paradox”: If we live in a static and eternal universe uniformly filled with stars, then the sky at night must be infinitely bright. The same observation had been made already earlier, by T. Digges (≈ 1580), by J. Kepler (1610) and by J.-P. de Cheseaux (1744).
- 1915** A. Einstein publishes the field equation of general relativity, in the version without a cosmological constant.
- 1917** A. Einstein introduces the cosmological constant in order to get static cosmological solutions. He finds a static dust solution with a positive cosmological constant, known as *Einstein’s static universe*.
- 1917** W. de Sitter finds an alternative cosmological world model which is a solution to Einstein’s vacuum field equation with a positive cosmological constant, known as the *de Sitter universe*.
- 1922/24** A. Friedmann finds the class of homogeneous and isotropic dust solutions to Einstein’s field equation (with or without a cosmological constant) named after him. A. Einstein opposes the idea of an expanding universe.
- 1927** G. Lemaître generalises the Friedmann solutions by allowing for a non-zero pressure. He discovers a linear relation between distance and redshift of galaxies, now known as the *Hubble law*, and interprets this as evidence for an expansion of our universe. He clearly expresses the idea that the universe began with an initial singularity which he called the “primeval atom”. In the 1960s, F. Hoyle coins the term “big bang”. The 1927 paper by Lemaître is written in French and remains unknown for a few years before A. Eddington arranges for an English translation in 1931.
- 1929** E. Hubble (re-)discovers the linear distance-redshift relation which becomes known as the *Hubble law*. This is soon recognised as observational evidence for an expanding universe. Hubble himself remains reluctant to accept this interpretation.
- 1932** A. Einstein is now convinced that his static universe from 1917 is not the correct world model. He abandons the idea of a cosmological constant and establishes together with W. de Sitter a spatially flat dust solution to the field equation. This *Einstein-deSitter universe* remains the favoured world model until the late 1990s when the accelerated expansion of the universe is discovered.
- 1936** F. Zwicky postulates the existence of dark matter in order to explain the stability of galaxy clusters.
- 1941** A. McKellar observes spectral lines from rotational transitions in cyanogen (CN) molecules in the interstellar medium. He comes to the conclusion that the interstellar medium must have a temperature of approximately 2.3 Kelvin. In hindsight, this is the first detection of the cosmic background radiation.
- 1946-49** G. Gamow and his PhD student R. Alpher develop a theory how hydrogen, helium and the heavier elements were created in the correct proportions after the initial singularity from a state they called “Ylem”. As a joke, Gamow puts H. Bethe, who actually was not involved, as a co-author on their paper (Alpher-Bethe-Gamow = $\alpha\beta\gamma$).

- 1948** R. Alpher and R. Herman predict the cosmic background radiation.
- 1948** H. Bondi, T. Gold and F. Hoyle invent the steady-state theory in which the universe is expanding, but the matter density remains constant because of a continuous creation of matter. The steady-state theory remains an important rival to the big-bang theory until the detection of the cosmic background radiation is recognised.
- 1956** W. Rindler introduces in his PhD Thesis the notions of event horizons and particle horizons.
- 1955-57** E. Leroux, T. Shmaonov and E. Ohm independently observe microwave radiation with the features of the predicted cosmic background radiation. Their observations remain widely unnoticed and are not recognised at the time.
- 1960-1970** V. Belinsky, I. Khalatnikov and E. Lifshits study general features of cosmological models with an initial singularity.
- 1964** A. Doroshkevich and I. Novikov make precise predictions for the existence of the cosmic background radiation.
- 1963-1969** R. Penrose and S. Hawking prove a series of theorems to the effect that the formation of a singularity is generic for solutions to Einstein's field equation where the energy-momentum tensor satisfies certain "energy conditions".
- 1964** A. Penzias and R. Wilson, when testing a new radio antenna, discover a mysterious isotropic noise. R. Dicke explains to them that they have found the predicted cosmic background radiation. Penzias and Wilson win the Nobel Prize in 1978.
- 1970-1980** V. Rubin performs a long series of spectral measurements to determine the rotation curves in galaxies. This provides strong evidence for the existence of dark matter.
- 1980** A. Starobinsky and A. Guth independently introduce the idea of inflation, i.e., that at an early stage the universe was exponentially expanding.
- 1989-1993** The satellite COBE investigates the cosmic background radiation. It is found that the radiation has a black-body spectrum with a temperature that is almost but not exactly isotropic. For these discoveries J. Mather and G. Smoot win the Nobel Prize in 2006.
- 1998** By using supernovae of type Ia as standard candles two teams independently find evidence for the fact that the expansion of the universe is accelerating. For the mysterious type of "matter" that causes the accelerated expansion M. Turner coins the word "dark energy". S. Perlmutter, B. Schmidt and A. Riess win the Nobel Prize in 2011 for the discovery of the accelerated expansion.
- 2001-2010** The satellite WMAP investigates the cosmic background radiation.
- 2008-2013** The satellite Planck complements these observations.

2 Brief review of general relativity

A general-relativistic spacetime is a pair (M, g) where:

- M is a four-dimensional manifold; local coordinates will be denoted (x^0, x^1, x^2, x^3) and Einstein's summation convention will be used for greek indices $\mu, \nu, \sigma, \dots = 0, 1, 2, 3$, for lower case latin indices $i, j, k, \dots = 1, 2, 3$ and for upper case latin indices $A, B, C, \dots = 1, 2$.
- g is a Lorentzian metric on M , i.e. a covariant second-rank tensor field, $g = g_{\mu\nu} dx^\mu \otimes dx^\nu$, that is
 - (a) symmetric, $g_{\mu\nu} = g_{\nu\mu}$, and
 - (b) non-degenerate with Lorentzian signature, i.e., for any $p \in M$ there are coordinates defined near p such that $g|_p = -(dx^0)^2 + (dx^1)^2 + (dx^2)^2 + (dx^3)^2$.

As the metric is non-degenerate, we may introduce contravariant metric components by

$$g^{\mu\nu} g_{\nu\sigma} = \delta_\sigma^\mu. \quad (1)$$

Here and in the following, δ_σ^μ denotes the Kronecker delta, $\delta_\sigma^\mu = 1$ if $\mu = \sigma$ and $\delta_\sigma^\mu = 0$ if $\mu \neq \sigma$. We use $g^{\mu\nu}$ and $g_{\sigma\tau}$ for raising and lowering indices, e.g.

$$g_{\rho\tau} A^\tau = A_\rho, \quad B_{\mu\nu} g^{\nu\tau} = B_\mu{}^\tau. \quad (2)$$

The metric contains all information about the spacetime geometry and thus about the gravitational field. In particular, the metric determines the following.

- The causal structure of spacetime:

A curve $s \mapsto x(s) = (x^0(s), x^1(s), x^2(s), x^3(s))$ is called

$$\left. \begin{array}{l} \text{spacelike} \\ \text{lightlike} \\ \text{timelike} \end{array} \right\} \iff g_{\mu\nu}(x(s)) \dot{x}^\mu(s) \dot{x}^\nu(s) \left\{ \begin{array}{l} > 0 \\ = 0 \\ < 0 \end{array} \right.$$

Timelike curves describe motion at subluminal speed and lightlike curves describe motion at the speed of light. Spacelike curves describe motion at superluminal speed which is forbidden for signals.

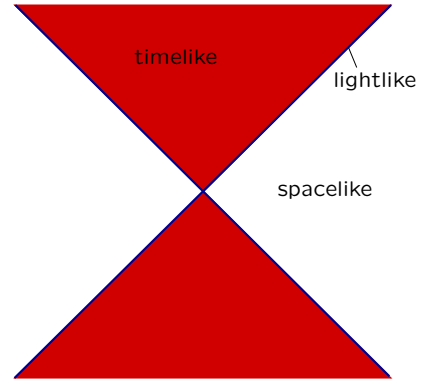


Figure 1: Light cone on the tangent space

For timelike curves we can choose the parametrisation such that $g_{\mu\nu}(x(\tau)) \dot{x}^\mu(\tau) \dot{x}^\nu(\tau) = -c^2$. The parameter τ is then called *proper time*.

The motion of a material continuum, e.g. of a fluid, can be described by a vector field $U = U^\mu \partial_\mu$ with $g_{\mu\nu} U^\mu U^\nu = -c^2$. The integral curves of U are to be interpreted as the worldlines of the fluid elements.

- The geodesics:

By definition, the geodesics are the solutions to the Euler-Lagrange equations

$$\frac{d}{ds} \frac{\partial \mathcal{L}(x, \dot{x})}{\partial \dot{x}^\mu} - \frac{\partial \mathcal{L}(x, \dot{x})}{\partial x^\mu} = 0 \quad (3)$$

of the Lagrangian

$$\mathcal{L}(x, \dot{x}) = \frac{1}{2} g_{\mu\nu}(x) \dot{x}^\mu \dot{x}^\nu . \quad (4)$$

These Euler-Lagrange equations take the form

$$\ddot{x}^\mu + \Gamma^\mu_{\nu\sigma}(x) \dot{x}^\nu \dot{x}^\sigma = 0 \quad (5)$$

where

$$\Gamma^\mu_{\nu\sigma} = \frac{1}{2} g^{\mu\tau} (\partial_\nu g_{\tau\sigma} + \partial_\sigma g_{\tau\nu} - \partial_\tau g_{\nu\sigma}) \quad (6)$$

are the so-called Christoffel symbols.

The Lagrangian $\mathcal{L}(x, \dot{x})$ is constant along a geodesic (see Worksheet 1), so we can speak of timelike, lightlike and spacelike geodesics. Timelike geodesics ($\mathcal{L} < 0$) are to be interpreted as the worldlines of freely falling particles, and lightlike geodesics ($\mathcal{L} = 0$) are to be interpreted as light rays.

The Christoffel symbols define a *covariant derivative* that takes tensor fields into tensor fields, e.g.

$$\nabla_\nu U^\mu = \partial_\nu U^\mu + \Gamma^\mu_{\nu\tau} U^\tau , \quad (7)$$

$$\nabla_\nu A_\mu = \partial_\nu A_\mu - \Gamma^\rho_{\nu\mu} A_\rho . \quad (8)$$

- The curvature.

The Riemannian curvature tensor is defined, in coordinate notation, by

$$R_{\mu\nu\sigma}{}^\tau = \partial_\mu \Gamma^\tau_{\nu\sigma} - \partial_\nu \Gamma^\tau_{\mu\sigma} + \Gamma^\tau_{\mu\rho} \Gamma^\rho_{\nu\sigma} - \Gamma^\tau_{\nu\rho} \Gamma^\rho_{\mu\sigma} . \quad (9)$$

The curvature tensor determines the relative motion of neighbouring geodesics: If $X = X^\mu \partial_\mu$ is a vector field whose integral curves are geodesics, and if $J = J^\nu \partial_\nu$ connects neighbouring integral curves of X (i.e., if the Lie bracket between X and J vanishes), then the *equation of geodesic deviation* or *Jacobi equation* holds:

$$(X^\mu \nabla_\mu)(X^\nu \nabla_\nu) J^\sigma = R_{\mu\nu\rho}{}^\sigma X^\mu J^\nu X^\rho . \quad (10)$$

If the integral curves of X are timelike, they can be interpreted as worldlines of freely falling particles. In this case the curvature term in the Jacobi equation gives the *tidal force* produced by the gravitational field.

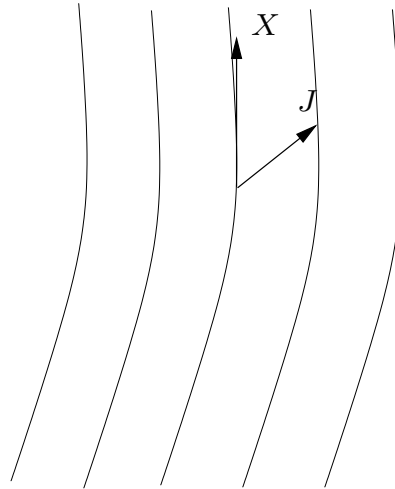


Figure 2: Jacobi equation

The curvature tensor satisfies the identities

$$R_{\mu\nu\sigma}{}^\tau = -R_{\nu\mu\sigma}{}^\tau, \quad (11)$$

$$R_{\mu\nu\sigma\tau} = -R_{\mu\nu\tau\sigma}, \quad (12)$$

$$R_{\mu\nu\sigma}{}^\tau + R_{\sigma\mu\nu}{}^\tau + R_{\nu\sigma\mu}{}^\tau = 0 \quad (1^{\text{st}} \text{ Bianchi identity}), \quad (13)$$

$$\nabla_\rho R_{\mu\nu\sigma}{}^\tau + \nabla_\nu R_{\rho\mu\sigma}{}^\tau + \nabla_\mu R_{\nu\rho\sigma}{}^\tau = 0 \quad (2^{\text{nd}} \text{ Bianchi identity}). \quad (14)$$

From the curvature tensor one defines the Ricci tensor

$$R_{\mu\nu} = R_{\sigma\mu\nu}{}^\sigma \quad (15)$$

and the Ricci scalar

$$R = R_{\mu\nu} g^{\mu\nu}. \quad (16)$$

The Ricci tensor is symmetric, $R_{\mu\nu} = R_{\nu\mu}$. In three dimensions, the curvature tensor is completely determined by the Ricci tensor and the metric tensor. In two dimensions, the curvature tensor is completely determined by the Ricci scalar and the metric tensor.

The spacetime metric is determined, in terms of its sources, by *Einstein's field equation*

$$R_{\mu\nu} - \frac{R}{2} g_{\mu\nu} + \Lambda g_{\mu\nu} = \kappa T_{\mu\nu}. \quad (17)$$

The curvature quantity

$$G_{\mu\nu} = R_{\mu\nu} - \frac{R}{2} g_{\mu\nu} \quad (18)$$

is called the *Einstein tensor field*, Λ is called the *cosmological constant*, and κ is called *Einstein's gravitational constant*. $T_{\mu\nu}$ is the *energy-momentum tensor* which describes the energy content of the spacetime. Examples for energy-momentum tensors will be given below.

Based on cosmological observations we believe that we live in a universe with a positive cosmological constant that is of the order of $\Lambda \approx 10^{-52} \text{m}^{-2}$, as we will discuss in detail later.

Einstein's gravitational constant is related to Newton's gravitational constant G according to $\kappa = 8\pi G/c^2$ as follows from the Newtonian limit of Einstein's theory.

By applying the operator ∇^μ to both sides of Einstein's field equation (17) one finds, with the help of the second Bianchi identity (14) and the rule $\nabla^\mu g_{\mu\nu} = 0$ which follows from the definition of the covariant derivative, that the left-hand side equals zero. Hence, Einstein's field equation can hold only if the energy-momentum tensor has a vanishing covariant divergence,

$$\nabla^\mu T_{\mu\nu} = 0. \quad (19)$$

This is interpreted as the law of conservation of energy in infinitesimally small spacetime regions. D. Lovelock has proven the following theorem in 1972: The only second-rank tensor fields with vanishing covariant divergence that can be formed out of the metric, its first and its second derivatives, are the tensor fields $\kappa^{-1}(R_{\mu\nu} - Rg_{\mu\nu}/2 + \Lambda g_{\mu\nu})$ where κ and Λ are constants. This remarkable theorem, which holds only in four dimensions, demonstrates that one has to give up either energy conservation in infinitesimally small regions or the idea that the field equation

should not be of higher than second order if one wants to set up an alternative to Einstein's theory in which the gravitational field is still determined by the metric tensor alone (i.e., no additional scalar fields, no torsion etc.).

The specific form of the energy-momentum tensor $T_{\mu\nu}$ depends on the matter model that is used for the source of the gravitational field. The most important cases are the following.

- Vacuum: $T_{\mu\nu} = 0$

Then the field equation simplifies to $R_{\mu\nu} = \Lambda g_{\mu\nu}$, as can be verified by calculating the trace of the field equation and then re-inserting the result into the field equation. The vacuum field equation is a system of ten scalar second-order non-linear partial differential equations for the ten independent metric coefficients $g_{\mu\nu}$. The best known solutions to Einstein's vacuum field equation with $\Lambda = 0$ are the Schwarzschild solution and the Kerr solution. An important solution with $\Lambda > 0$ is the deSitter metric and with $\Lambda < 0$ the anti-deSitter metric. Both will be discussed in this course.

- Electrovacuum: $T_{\mu\nu} = F_{\mu\alpha}F_{\nu}^{\alpha} - \frac{1}{4}g_{\mu\nu}F_{\alpha\beta}F^{\alpha\beta}$

In this case Einstein's field equation together with Maxwell's equations gives a system of partial differential equations for the $g_{\mu\nu}$ and the electromagnetic field strength $F_{\mu\nu}$. The first Maxwell equation $\nabla_{\mu}F_{\nu\sigma} + \nabla_{\nu}F_{\sigma\mu} + \nabla_{\sigma}F_{\mu\nu} = 0$ is automatically satisfied if we assume that $F_{\mu\nu}$ derives from a potential, $F_{\mu\nu} = \nabla_{\mu}A_{\nu} - \nabla_{\nu}A_{\mu}$, and the second Maxwell equation $\nabla^{\mu}F_{\mu\nu} = 0$ is equivalent to the conservation law $\nabla^{\mu}T_{\mu\nu} = 0$ (which is a consequence of Einstein's field equation). The best-known electrovacuum solutions without a cosmological constant are the Reissner-Nordström solution (field outside of a charged spherically symmetric static object) and the Kerr-Newman solution (field of a charged and rotating black hole).

- Perfect fluid: $T_{\mu\nu} = \left(\varepsilon + p\right)\frac{1}{c^2}U_{\mu}U_{\nu} + pg_{\mu\nu}$

For solving Einstein's field equation with a perfect-fluid source one has to specify an equation of state linking the pressure p to the energy density ε . In cosmology one usually considers equations of state of the form $p = w\varepsilon$ with a constant w . The simplest example is a "dust", $w = 0$. If the equation of state has been fixed, Einstein's equation together with the Euler equation

$$\left(\varepsilon + p\right)\frac{1}{c^2}U^{\rho}\nabla_{\rho}U^{\sigma} + \nabla_{\tau}p\left(g^{\tau\sigma} + \frac{1}{c^2}U^{\tau}U^{\sigma}\right) = 0 \quad (20)$$

gives a system of partial differential equations for the $g_{\mu\nu}$, the four-velocity U^{ρ} and the energy density ε . Actually, the Euler equation is a consequence of the energy conservation law $\nabla_{\mu}T^{\mu\nu} = 0$ which follows from Einstein's field equation. So in the case of a perfect fluid Einstein's field equation determines the equation of motion of the matter source.

Perfect fluid solutions without a cosmological constant are of interest as models for the interior of stars. The interior Schwarzschild solution is an example; it describes a spherically symmetric static star with constant density ε . In this course we will intensively study the Friedmann-Lemaître solutions, which are the simplest cosmological models of our universe. They are perfect fluid solutions, possibly with a cosmological constant. We will also consider some cosmological models that are homogeneous but not isotropic (Bianchi models).

- Scalar field: $T_{\mu\nu} = \nabla_\mu \phi \nabla_\nu \phi - \left(\frac{1}{2} \nabla_\rho \phi \nabla^\rho \phi - V(\phi) \right) g_{\mu\nu}$

Here ϕ is a real scalar field and $V(\phi)$ is a potential. Einstein's equation together with the equation $\nabla^\mu T_{\mu\nu} = 0$ (which is a consequence of the Einstein equation) give a system of partial differential equations for $g_{\mu\nu}$ and ϕ . For a quadratic potential $V(\phi) = (m^2/2)\phi^2$ the scalar field equation $\nabla^\mu T_{\mu\nu} = 0$ gives the Klein-Gordon equation $g^{\mu\nu} \nabla_\mu \nabla_\nu \phi - m^2 \phi = 0$. In cosmology one usually considers potentials that contain higher-than-quadratic terms in ϕ which describe a self-interaction of the scalar field. The most important examples of scalar fields in cosmology are the *inflaton* field that drives inflation and the *cosmon* or *quintessence* fields that are ways of modelling *dark energy*. We will discuss them in some detail. Also the *Higgs* field plays a role in cosmology.

3 Homogeneous and isotropic cosmology

In Section 3.1 we give a characterisation of spacetime models that are spatially homogeneous and isotropic and we discuss their properties. This consideration is purely kinematic, i.e., Einstein's field equation is not used. In the subsequent section we will then solve the field equation within the class of spatially homogeneous and isotropic spacetimes, and we will discuss some solutions and their properties in detail.

3.1 Robertson-Walker spacetimes

By definition, a Robertson-Walker spacetime is a general-relativistic spacetime that can be sliced into 3-dimensional space-like submanifolds that are homogeneous and isotropic, see Figure 3. The general form of such metrics was determined independently by H. P. Robertson and by A. Walker in 1935/37.

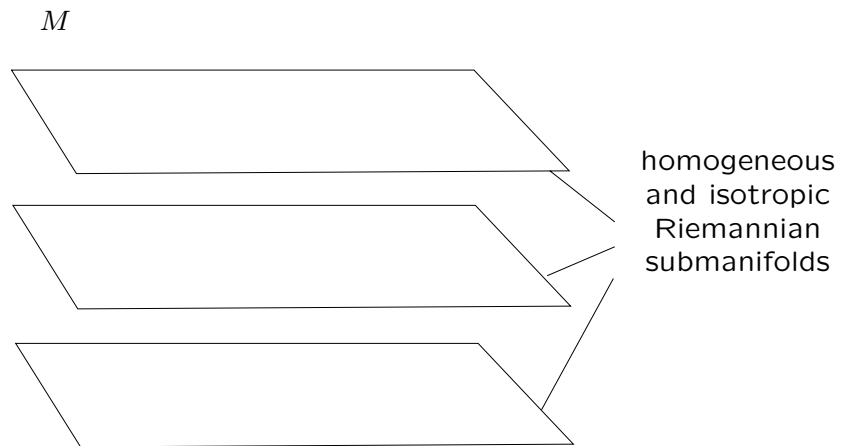


Figure 3: Robertson-Walker spacetime

The first step is to determine the geometry of the time slices. The assumption that they are spacelike means that they inherit from the spacetime metric a Riemannian (i.e., positive definite) metric. So the task is to determine all 3-dimensional Riemannian manifolds that are homogeneous and isotropic. Here “homogeneous” means that there are no distinguished points and “isotropic” means that there are no distinguished directions.

We consider a 3-dimensional manifold with a Riemannian metric $g_{ik}dx^i \otimes dx^k$ whose curvature tensor we denote by $R_{ijk}{}^l$. (Recall our convention of having latin indices running from 1 to 3.) As $R_{ij}{}^{kl} = -R_{ji}{}^{kl}$ and $R_{ij}{}^{kl} = -R_{ij}{}^{lk}$, we can define at each point a linear map from the space of antisymmetric second-rank tensors $\Lambda_2 = \{\omega_{li}dx^l \otimes dx^i | \omega_{li} = -\omega_{il}\}$ onto itself by

$$\Lambda_2 \longrightarrow \Lambda_2 \quad (21)$$

$$\omega_{kl}dx^k \otimes dx^l \longmapsto \hat{\omega}_{ij}dx^i \otimes dx^j = R_{ij}{}^{kl}\omega_{kl}dx^i \otimes dx^j.$$

Owing to the first Bianchi identity, this linear map is symmetric (with respect to the positive definite scalar product induced by the metric), so it has three linearly independent eigenvectors. If the eigenvalues would be different from each other, the eigenvectors would define distinguished directions in the tangent space, in contradiction to the assumption of isotropy. So the three eigenvalues must be equal, i.e., the linear map must be a multiple of the identity map,

$$R_{ij}{}^{kl} = K(\delta_i^k \delta_j^l - \delta_i^l \delta_j^k) \quad (22)$$

with a scalar factor K . (Note that $(\delta_i^k \delta_j^l - \delta_i^l \delta_j^k)\omega_{kl} = \omega_{ij} - \omega_{ji} = 2\omega_{ij}$, i.e., that the antisymmetrised product of Kroneckers acts, indeed, as the identity operator on antisymmetric tensor fields, up to a factor of 2 that was absorbed in the K .) The condition of homogeneity requires K to be a constant. It is common to write (22) in terms of the covariant components of the curvature tensor. Then the conditions of homogeneity and isotropy require that

$$R_{ijkl} = K(g_{ik}g_{jl} - g_{il}g_{jk}), \quad K = \text{constant}.$$

One says that a Riemannian manifold is a “space of constant curvature” if this condition holds. It is interesting to note that in all dimensions $n > 2$, and thus in particular in the case $n = 3$ considered here, K is necessarily a constant if the condition of isotropy holds at every point, i.e., the condition of homogeneity need not be required separately. To prove this, it is sufficient to apply the second Bianchi identity to the curvature tensor from (22) and to contract two pairs of indices away; this results in $(n-1)(n-2)\nabla_i K = 0$.

We have demonstrated that homogeneous and isotropic Riemannian manifolds are necessarily spaces of constant curvature. One can prove that, conversely, the spaces of constant curvature are precisely those Riemannian manifolds for which the conditions of homogeneity and isotropy hold *locally*. More precisely, one can prove the following:

- Any two Riemannian manifolds of constant curvature with the same dimension and the same K are locally isometric. In other words, locally there is only one such geometry for each K . The global structure is not uniquely determined. Spaces of constant curvature that are geodesically complete are called *space forms*. For low dimensions, the space forms have been classified.

- A space of constant curvature is a space with the maximal number of local symmetries. Local symmetries are characterised in terms of *Killing vector fields*, see Worksheet 2. By definition, a vector field $K = K^i \partial_i$ is a Killing vector field if the Lie derivative of the metric in the direction of K vanishes. If K has no zeros, this is equivalent to saying that there is a coordinate system in which $K = \partial_1$ and the metric coefficients are independent of x^1 . One can show that the linear combination of two Killing vector fields with *constant* coefficients is again a Killing vector field, i.e., that the Killing vector fields form a vector space over the real numbers. One can further show that the dimension of this vector space for an n -dimensional manifold cannot be bigger than $n(n+1)/2$. So on a 3-dimensional manifold there are at most six linearly independent Killing vector fields. The maximal number is just reached in the case of homogeneity and isotropy where we have 3 translations and 3 rotations.

So the time-slices in a Robertson-Walker spacetime are 3-dimensional Riemannian manifolds of constant curvature. Before turning to the 3-dimensional case we first consider the 2-dimensional Riemannian manifolds of constant curvature which are easier visualised. In this case we use capital indices, taking the values 1 and 2, and we write the curvature tensor as

$$R_{ABCD} = K(g_{AC}g_{BD} - g_{AD}g_{BC}), \quad K = \text{constant}. \quad (23)$$

This implies that the Ricci tensor is

$$R_{BC} = R_{ABC}{}^A = K(g_{BC} - 2g_{BC}) = -K g_{BC} \quad (24)$$

and the Ricci scalar is

$$R = -2K. \quad (25)$$

So with our conventions the Ricci scalar is negative for a space of positive constant curvature and vice versa. We consider the cases $K = 0$, $K > 0$ and $K < 0$ separately.

$K = 0$: The condition $K = 0$ means that $R_{ABCD} = 0$ which is certainly true for the Euclidean plane.

$$g = dx^2 + dy^2. \quad (26)$$

If we transform to polar coordinates,

$$x = a \chi \cos \varphi, \quad (27)$$

$$y = a \chi \sin \varphi, \quad (28)$$

where a is a constant with the dimension of a length and χ is a dimensionless radial coordinate, the metric reads

$$g = a^2(d\chi^2 + \chi^2 d\varphi^2). \quad (29)$$

All other 2-dimensional Riemannian manifolds of constant curvature $K = 0$ are locally isometric to the Euclidean plane. The space forms, i.e., the geodesically complete cases, can be constructed as quotient manifolds from the plane, i.e., by identifying points on the plane. In addition to the plane itself, there are four of them: The cylinder, the torus, the Moebius strip and the Klein bottle.

$K > 0$: An obvious candidate for a space of constant positive curvature is the sphere of radius a which is defined as a submanifold of Euclidean 3-space by the equation $X^2 + Y^2 + Z^2 = a^2$. We can parametrise the sphere by angle coordinates (χ, φ) via

$$X = a \cos \varphi \sin \chi, \quad (30)$$

$$Y = a \sin \varphi \sin \chi, \quad (31)$$

$$Z = a \cos \chi. \quad (32)$$

Then we have on the sphere

$$dX = a \cos \varphi \cos \chi d\chi - a \sin \varphi \sin \chi d\varphi, \quad (33)$$

$$dY = a \sin \varphi \cos \chi d\chi + a \cos \varphi \sin \chi d\varphi, \quad (34)$$

$$dZ = -a \sin \chi d\chi. \quad (35)$$

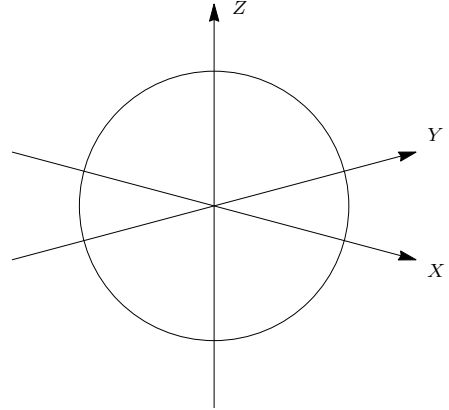


Figure 4: Sphere

Inserting these results into the expression $dX^2 + dY^2 + dZ^2$ demonstrates that the Euclidean 3-metric induces on the sphere the metric

$$g = a^2(d\chi^2 + \sin^2 \chi d\varphi^2). \quad (36)$$

By calculating the Christoffel symbols and, thereupon, the Riemann tensor one easily verifies that the condition of constant curvature is indeed satisfied with $K = 1/a^2$.

Again, all other 2-dimensional spaces of constant curvature $K > 0$ are locally isometric to the sphere. The only other space form is 2-dimensional projective space which results from the sphere by identifying antipodal points. 2-dimensional projective space cannot be globally embedded into Euclidean 3-space.

$K < 0$: Guided by the case $K > 0$, one tries a similar construction but now in a way that the curvature comes out as $K = -1/a^2$. This requires to change some signs in the signature of the ambient space and of the embedding formula: The manifold is now given by the equation $X^2 + Y^2 - T^2 = -a^2$ and the ambient space has Minkowskian signature, i.e., the metric is $dX^2 + dY^2 - dT^2$. This defines a hyperboloid which can be parametrised as

$$X = a \cos \varphi \sinh \chi, \quad (37)$$

$$Y = a \sin \varphi \sinh \chi, \quad (38)$$

$$T = a \cosh \chi. \quad (39)$$

Then we have on the hyperboloid

$$dX = a \cos \varphi \cosh \chi d\chi - a \sin \varphi \sinh \chi d\varphi, \quad (40)$$

$$dY = a \sin \varphi \cosh \chi d\chi + a \cos \varphi \sinh \chi d\varphi, \quad (41)$$

$$dT = a \sinh \chi d\chi. \quad (42)$$

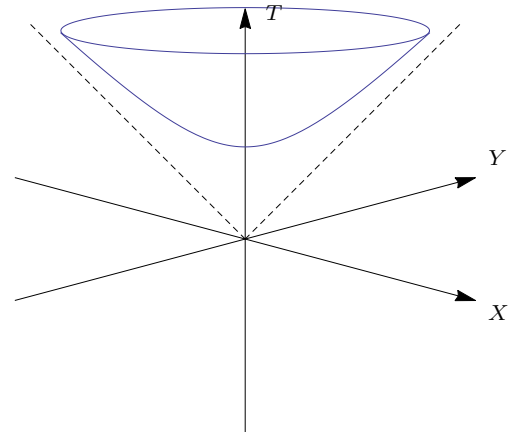


Figure 5: Hyperboloid

Inserting these results into the expression $dX^2 + dY^2 - dZ^2$ demonstrates that the 3-dimensional Minkowski metric induces on the hyperboloid the metric

$$g = a^2(d\chi^2 + \sinh^2\chi d\varphi^2). \quad (43)$$

Again, the Christoffel symbols and the components of the curvature tensor are readily calculated and one finds that, indeed, this is a space of constant curvature with $K = -1/a^2$.

The space of constant negative curvature is known as *hyperbolic space*, as *Lobachevsky space* or as *Lobachevsky-Bolyai space*, named after Nikolai Lobachevsky and János Bolyai who independently discovered this geometry in the 1820s. In hyperbolic space, the sum of the angles in a geodesic triangle is smaller than π while on a sphere it is bigger. Hyperbolic geometry satisfies all axioms of Euclid, i.e., all axioms of flat geometry, with the exception of the parallel axiom.

In addition to the isometric embedding into 3-dimensional Minkowski space, hyperbolic space can be represented in various different ways. However, it cannot be isometrically embedded into 3-dimensional Euclidean space. (A non-global embedding is possible in the form of a surface of revolution generated by a *tractrix*.)

Other space forms of negative curvature can be constructed as quotient manifolds from the hyperboloid in Minkowski spacetime.

We summarise our results in the following way: A 2-dimensional Riemannian manifold of constant curvature K is given by the metric

$$g = a^2(d\chi^2 + \xi(\chi)^2 d\varphi^2), \quad \xi(\chi) = \begin{cases} \sin \chi & \text{for } K = \frac{1}{a^2} > 0, \\ \chi & \text{for } K = 0, \\ \sinh \chi & \text{for } K = -\frac{1}{a^2} < 0. \end{cases} \quad (44)$$

The transition from 2 to 3 dimensions simply requires the circle parametrised by φ to be replaced with a sphere parametrised by standard spherical coordinates (ϑ, φ) , i.e., a 3-dimensional Riemannian manifold of constant curvature K is given by the metric

$$g = a^2(d\chi^2 + \xi(\chi)^2\{d\vartheta^2 + \sin^2\vartheta d\varphi^2\}), \quad \xi(\chi) = \begin{cases} \sin \chi & \text{for } K = \frac{1}{a^2} > 0, \\ \chi & \text{for } K = 0, \\ \sinh \chi & \text{for } K = -\frac{1}{a^2} < 0. \end{cases} \quad (45)$$

One cannot visualise a 3-dimensional space of constant curvature, except in the case $K = 0$, but one can visualise the equatorial section $\vartheta = \pi/2$ which is a 2-dimensional space of constant curvature.

There are two alternative coordinate representations of spaces of constant curvature. Firstly, we can make a coordinate transformation $(\chi, \vartheta, \varphi) \mapsto (\xi, \vartheta, \varphi)$ with $\xi(\chi)$ from (45). For $K > 0$ we find

$$\xi = \sin \chi, \quad d\xi = \cos \chi d\chi = \sqrt{1 - \xi^2} d\chi, \quad (46)$$

for $K = 0$ we simply have

$$\xi = \chi, \quad d\xi = d\chi \quad (47)$$

and for $K < 0$

$$\xi = \sinh \chi, \quad d\xi = \cosh \chi d\chi = \sqrt{1 + \xi^2} d\chi. \quad (48)$$

The three cases can be written in a unified way as

$$d\chi = \frac{d\xi}{\sqrt{1 - k\xi^2}} \quad \text{where} \quad \begin{cases} k = 1 & \text{if } K > 0, \\ k = 0 & \text{if } K = 0, \\ k = -1 & \text{if } K < 0, \end{cases} \quad (49)$$

so the metric becomes

$$g = a^2 \left(\frac{d\xi^2}{1 - k\xi^2} + \xi^2 \{ d\vartheta^2 + \sin^2 \vartheta d\varphi^2 \} \right) \quad (50)$$

where $k = 1, 0$ or -1 .

Secondly, another coordinate transformation $(\xi, \vartheta, \varphi) \mapsto (\rho, \vartheta, \varphi)$ can be found such that the metric takes the form

$$g = \frac{a^2 \left(d\rho^2 + \rho^2 \{ d\vartheta^2 + \sin^2 \vartheta d\varphi^2 \} \right)}{\left(1 + \frac{k}{4} \rho^2 \right)^2}. \quad (51)$$

This transformation and the geometric meaning of the coordinate ρ will be discussed in Worksheet 2. All three coordinate representations of spaces of constant curvature are frequently used for applications in cosmology.

Now we know the 3-dimensional homogeneous and isotropic Riemannian manifolds. This was the first step for constructing Robertson-Walker spacetimes. The second step is adding the time dimension.

As we assume that the spacetime admits a slicing into 3-dimensional Riemannian submanifolds of constant curvature, there is a distinguished timelike direction at each point, viz., the direction perpendicular to the slices. These timelike directions have as their integral curves a distinguished family of observer worldlines. Along these worldlines we can use proper time as a parametrisation. Homogeneity requires that the proper time that elapses between any two fixed slices is the same along all worldlines perpendicular to the slices. Hence, the proper time parametrisation defines a time coordinate t on the spacetime such that the slices become submanifolds $\{t = \text{constant}\}$.

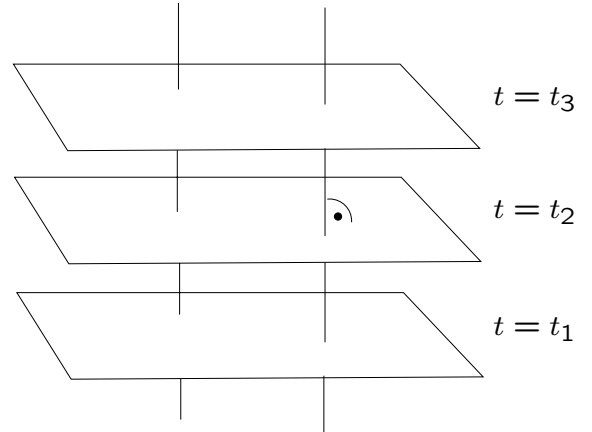


Figure 6: Robertson-Walker spacetime

Then the metric reads

$$g = -c^2 dt^2 + a(t)^2 (d\chi^2 + \xi(\chi)^2 \{d\vartheta^2 + \sin^2 \vartheta d\varphi^2\}) = \quad (52)$$

$$= -c^2 dt^2 + a(t)^2 \left(\frac{d\xi^2}{1 - k\xi^2} + \xi^2 \{d\vartheta^2 + \sin^2 \vartheta d\varphi^2\} \right) = \quad (53)$$

$$= -c^2 dt^2 + \frac{a(t)^2 (d\rho^2 + \rho^2 \{d\vartheta^2 + \sin^2 \vartheta d\varphi^2\})}{\left(1 + \frac{k}{4} \rho^2\right)^2} \quad (54)$$

which is the general form of a Robertson-Walker metric. We have no mixed metric components, $g_{0i} = 0$, because the t -lines are perpendicular to the $\{t = \text{constant}\}$ -slices. The coefficient in front of dt^2 must be $-c^2$ because t is supposed to be proper time on the worldlines perpendicular to the slices. And the spatial part is a metric of constant curvature with a scale factor that may depend on t , indicating that the universe may expand or contract.

A Robertson-Walker universe is locally (but not globally) fixed by the curvature parameter k and by the scale factor $a(t)$. There is a freedom in choosing the topology. We say that for $k = 1$ the *natural topology* is a 3-sphere while for $k = 0$ and $k = -1$ it is \mathbb{R}^3 . However, we are free to form quotient manifolds, e.g., to consider a Robertson-Walker universe with $k = 0$ that has a toroidal spatial topology. Therefore it is misleading to call the universes with $k > 0$ “closed” and the ones with $k \leq 0$ “open”. By the same token, the time slices of a Robertson-Walker universe may have a finite volume (and be geodesically complete) even if $k = 0$ or $k = -1$.

The t -lines, i.e., the timelike curves perpendicular to the slices of constant curvature, are called the worldlines of the *standard observers*. In particular in expanding Robertson-Walker models it is also common to call the flow of the vector field ∂_t the *Hubble flow*.

If the scale factor takes the value 0, the metric degenerates. Therefore the function a has to be restricted to a maximal interval on which $a(t) \neq 0$. As only $a(t)^2$ matters, we may choose $a(t) > 0$ without loss of generality. So a is a map of the form

$$a :] t_i, t_f [\longrightarrow] 0, \infty [\quad (55)$$

$$t \longmapsto a(t).$$

Here either $t_i = -\infty$ or t_i is finite and, analogously, either $t_f = \infty$ or t_f is finite. The suffix i stands for “initial” and the suffix f stands for “final”.

Models with t_i finite and $a(t) \rightarrow 0$ for $t \rightarrow t_i$ are called *big bang* models. Models with t_f finite and $a(t) \rightarrow 0$ for $t \rightarrow t_f$ are called *big crunch* models. A big bang is a singularity in the sense that all standard observers are compressed into zero volume at some finite proper time back into the past, and the same is true for a big crunch at some finite proper time into the future. Note, however, that this singularity is not necessarily a curvature singularity. The Milne model, see Figure 8 below, is a big bang model where the spacetime metric is perfectly regular at the big bang; it is the family of standard observers that becomes pathological.

Models with t_f finite and $a(t) \rightarrow \infty$ for $t \rightarrow t_f$ have also become of interest in recent years. They are called *big rip* models. For models with t_i finite and $a(t) \rightarrow \infty$ for $t \rightarrow t_i$ there is no special name.

A (trivial) example of a Robertson-Walker spacetime is Minkowski spacetime in inertial coordinates. In this case $k = 0$ and $a = \text{constant}$. Of course, from one such slicing of Minkowski spacetime into flat spacelike hyperplanes we can change to another one by a Lorentz boost, see the Figure 7. So this example demonstrates that the family of standard observers in a Robertson-Walker spacetime is not necessarily unique.

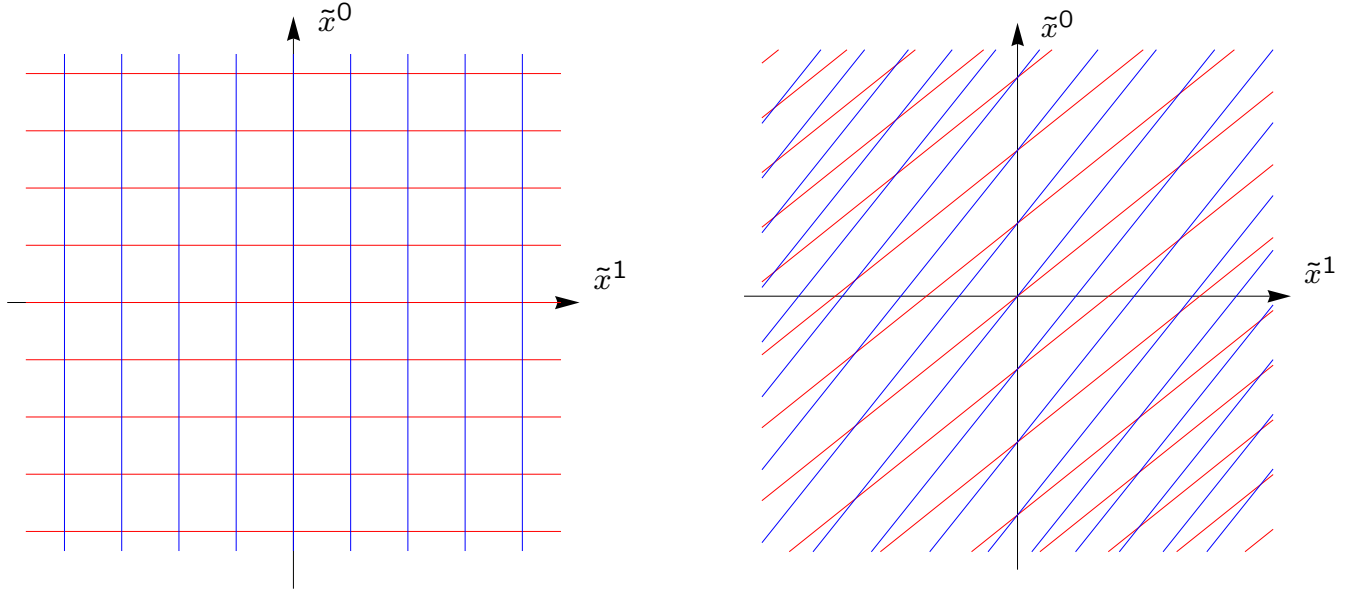


Figure 7: Slicings of Minkowski spacetime

Another (less trivial) example of a Robertson-Walker universe is the *Milne model* which was suggested, as a special-relativistic model of our universe, by E. Milne in 1935. It has hyperbolic spatial geometry, $k = -1$, and a scale factor $a(t) = ct$.

The Milne model can be isometrically embedded into Minkowski spacetime where it covers the future light-cone of an event, see Figure 8. Therefore, it is a vacuum solution of Einstein's field equation with $\Lambda = 0$ and thus not a realistic model of our universe. The slices $t = \text{constant}$ are 3-dimensional hyperboloids (red) and the worldlines of the standard observers are straight lines (blue). This model starts at $t_i = 0$ with a big bang in the sense that at this time all standard observers were compressed into one point, but this is of course neither a curvature singularity nor a point of infinite matter density: The curvature of Minkowski spacetime is zero and so is the matter density if Einstein's field equation is considered.

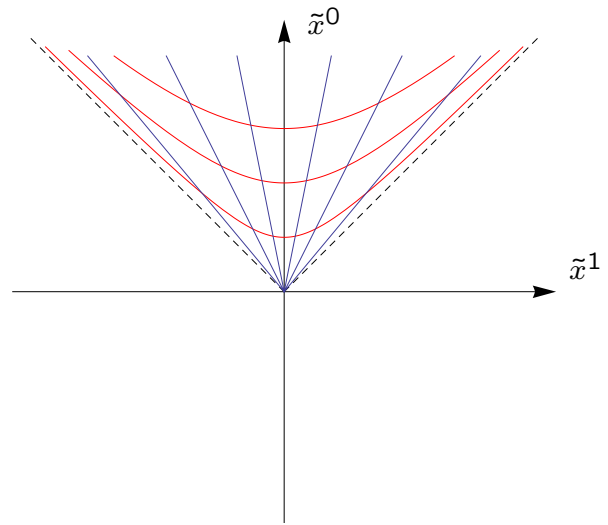


Figure 8: Milne model

The geodesics in a Robertson-Walker spacetime are the solutions to the Euler-Lagrange equations

$$\frac{d}{ds} \left(\frac{\partial \mathcal{L}(x, \dot{x})}{\partial \dot{x}^\mu} \right) = \frac{\partial \mathcal{L}(x, \dot{x})}{\partial x^\mu} \quad (56)$$

with the Lagrangian

$$\mathcal{L}(x, \dot{x}) = \frac{1}{2} g_{\mu\nu} \dot{x}^\mu \dot{x}^\nu = \frac{1}{2} \left(-c^2 \dot{t}^2 + a(t)^2 \left\{ \dot{\chi}^2 + \xi(\chi)^2 (\dot{\vartheta}^2 + \sin^2 \vartheta \dot{\varphi}^2) \right\} \right). \quad (57)$$

Here the overdot means derivative with respect to the affine parameter s . The four components of the Euler-Lagrange equation read

$$\frac{d}{ds} \left(-c^2 \dot{t} \right) = a(t) \frac{da(t)}{dt} \left\{ \dot{\chi}^2 + \xi(\chi)^2 (\dot{\vartheta}^2 + \sin^2 \vartheta \dot{\varphi}^2) \right\}, \quad (58)$$

$$\frac{d}{ds} \left(a(t)^2 \dot{\chi} \right) = a(t)^2 \xi(\chi) \frac{d\xi(\chi)}{d\chi} (\dot{\vartheta}^2 + \sin^2 \vartheta \dot{\varphi}^2), \quad (59)$$

$$\frac{d}{ds} \left(a(t)^2 \xi(\chi)^2 \dot{\vartheta} \right) = a(t)^2 \xi(\chi)^2 \sin \vartheta \cos \vartheta \dot{\varphi}^2, \quad (60)$$

$$\frac{d}{ds} \left(a(t)^2 \xi(\chi)^2 \sin^2 \vartheta \dot{\varphi} \right) = 0. \quad (61)$$

From (60) and (61) we read that, if we choose initial conditions $\dot{\vartheta}(s_o) = 0$ and $\dot{\varphi}(s_o) = 0$, then the solution satisfies $\dot{\vartheta}(s) = 0$ and $\dot{\varphi}(s) = 0$ for all s . In other words, a geodesic remains radial if it starts in the radial direction. Of course, this is an obvious consequence of the isotropy. Moreover, the spacetime is spatially homogeneous. Therefore, if we want to know all the geodesics issuing from a certain event, we may choose the coordinate system such that this event is on the worldline $\chi = 0$, i.e., at the spatial origin of the coordinate system. Combining these two observations tells us that it is sufficient to consider radial geodesics, $\dot{\vartheta} = 0$ and $\dot{\varphi} = 0$, with the initial condition $(t(s_o) = t_o, \chi(s_o) = 0)$. Then (60) and (61) are automatically satisfied. For analysing the remaining equations we recall that the Lagrangian is a constant of motion,

$$-c^2 \dot{t}^2 + a(t)^2 \dot{\chi}^2 = -\varepsilon c^2 \quad (62)$$

where $\varepsilon = 1$ for timelike geodesics and $\varepsilon = 0$ for lightlike ones. (For spacelike geodesics we have $\varepsilon = -1$ but we will not consider this case because it is not of interest in view of physics.) As $\dot{\vartheta} = 0$ and $\dot{\varphi} = 0$, equation (59) requires

$$\frac{d}{d\tau} \left(a(t)^2 \dot{\chi} \right) = 0, \quad a(t)^2 \dot{\chi} = A = \text{constant}. \quad (63)$$

The constant of motion A determines the initial velocity with respect to the standard observers. Equation (58) will not be needed in the following because it yields no additional information. Inserting (63) into (62) results in

$$\begin{aligned} -c^2 \frac{\dot{t}^2}{\dot{\chi}^2} + a(t)^2 &= \frac{-\varepsilon c^2 a(t)^4}{A^2}, & c^2 \left(\frac{dt}{d\chi} \right)^2 &= a(t)^2 \left(1 + \varepsilon \frac{c^2 a(t)^2}{A^2} \right), \\ d\chi &= \frac{\pm A c dt}{a(t) \sqrt{A^2 + \varepsilon c^2 a(t)^2}}. \end{aligned} \quad (64)$$

We consider now the case of timelike geodesics, $\varepsilon = 1$. If we integrate, for this case, (64) with the initial conditions $(t(\tau_o) = t_o, \chi(\tau_o) = 0)$, we find

$$\chi = \pm \int_{t_o}^t \frac{A c d\tilde{t}}{a(\tilde{t}) \sqrt{A^2 + c^2 a(\tilde{t})^2}}. \quad (65)$$

Note that χ takes only positive values. Without loss of generality we may assume that $A \geq 0$. Then we have to choose the plus sign for times $t > t_o$ and the minus sign for times $t < t_o$. For $A = 0$ one gets the t -line which we have chosen as the spatial origin. As we can choose any t -line, we have thus proven that all the standard observers are freely falling, i.e., that they stay on their worldlines without a thrust. Of course, this is an obvious consequence of the isotropy: If the standard observers were non-geodesic, they had a non-vanishing 4-acceleration and this 4-acceleration would distinguish a spatial direction.

Note that in a spatially compact universe χ cannot take all positive values. This happens, e.g., in a Robertson-Walker universe with $k = 1$ and the spatial topology of a 3-sphere, where χ is restricted to values between 0 and π . In this case a timelike geodesic with $A \neq 0$ may return to the same point in space (i.e., to the same standard observer) from where it started.

Next we evaluate (64) for lightlike geodesics, $\varepsilon = 0$. We find, upon integration with the chosen initial condition,

$$\chi = \pm \int_{t_o}^t \frac{c d\tilde{t}}{a(\tilde{t})}. \quad (66)$$

As before, the plus sign must be chosen for $t > t_o$ and the minus sign for $t < t_o$.

Figure 9 illustrates radial lightlike geodesics that issue into the future (blue) and radial lightlike geodesics that issue into the past (red) for a case where χ is restricted to an interval $0 < \chi < \chi_{\max}$ (for a spherical universe $\chi_{\max} = \pi$) and where the scale factor restricts the t coordinate to a finite interval $t_i < t < t_f$. Note that every point in this diagram represents a sphere because the ϑ and φ coordinates are not shown. In a non-spherical but spatially compact universe χ_{\max} may depend on ϑ and φ ; this happens, e.g., in a Robertson-Walker universe with $k = 0$ and a toroidal topology.

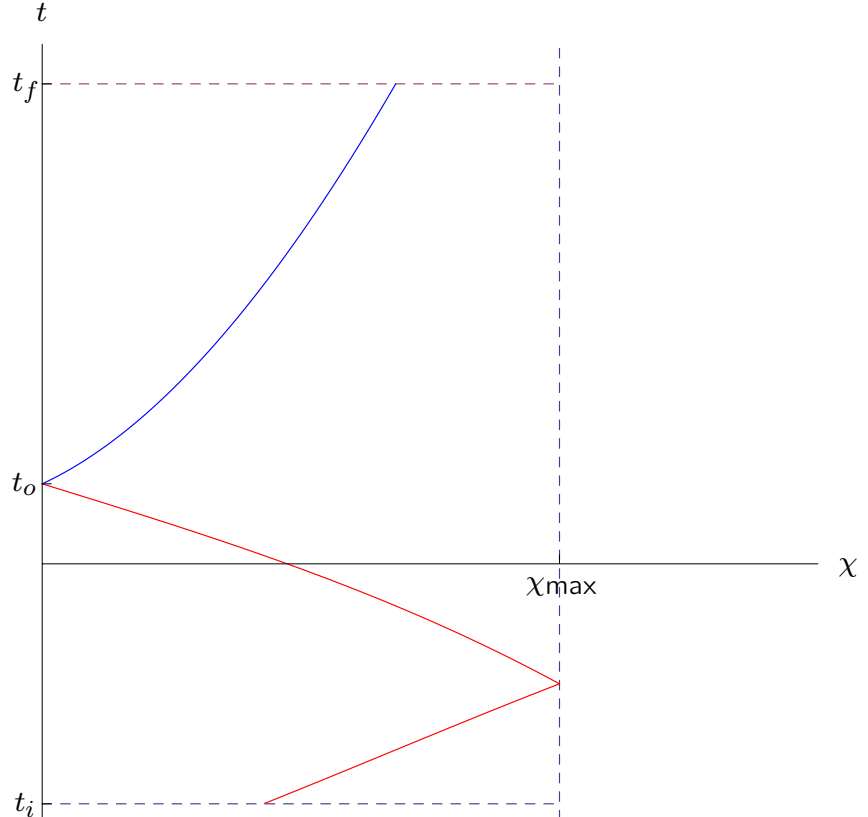


Figure 9: Lightlike geodesics in a Robertson-Walker spacetime

We will now discuss three observable features of Robertson-Walker universes all of which are related to lightlike geodesics: The redshift, the horizons and various distance measures.

- (i) The redshift: The redshift can be defined, for any pair of worldlines parametrised by proper time in any spacetime model, in the following way. Assume that from one of the worldlines light rays are emitted at proper times labelled t_e and that they are received on the other worldline at proper times labelled t_o . The indices e and o stand for “emitter” and “observer”, respectively. Then we can calculate the frequency ratio

$$\frac{dt_o}{dt_e} = \lim_{\Delta t_e \rightarrow 0} \frac{\Delta t_o}{\Delta t_e} = \frac{\omega_e}{\omega_o} = \frac{\lambda_o}{\lambda_e} \quad (67)$$

where we have used that a process with period Δt_e has frequency $\omega_e = 2\pi/\Delta t_e$ and similarly for Δt_o and ω_o . Also, for light propagating in vacuo we can use the dispersion relation $\omega\lambda = c$ to convert a frequency ω into a wave-length λ . The limit $\Delta t_e \rightarrow 0$ is necessary to make the result unique. Astronomers define the redshift z as the change of wave-length divided by the emitted wave-length, i.e.

$$z = \frac{\lambda_o - \lambda_e}{\lambda_e} = \frac{dt_o}{dt_e} - 1. \quad (68)$$

Figure 10 illustrates this situation for the case that both the emitter and the observer are standard observers in a Robertson-Walker spacetime. Without loss of generality, we choose the coordinate system such that the observer is at the origin, $\chi_o = 0$, while the emitter is at a certain radius coordinate χ_e . Here it is important to recall that the time coordinate t gives proper time for the standard observers, i.e., $\omega_e = 2\pi/\Delta t_e$ and $\omega_o = 2\pi/\Delta t_o$ are, indeed, the frequencies as measured with standard clocks.

Differentiating the equation for lightlike geodesics

$$\chi_e = - \int_{t_o}^{t_e} \frac{c dt}{a(t)} = \int_{t_e}^{t_o} \frac{c dt}{a(t)} \quad (69)$$

with respect to t_e yields

$$0 = \frac{c}{a(t_o)} \frac{dt_o}{dt_e} - \frac{c}{a(t_e)},$$

hence

$$\frac{dt_o}{dt_e} = \frac{a(t_o)}{a(t_e)}. \quad (70)$$

Inserting this result into (68) yields

$$1 + z = \frac{a(t_o)}{a(t_e)}. \quad (71)$$

This is the redshift law for standard observers in a Robertson-Walker universe. If the observer and/or the emitter is not a standard observer, one has to supply additional Doppler factors that are determined by the velocity relative to a standard observer.

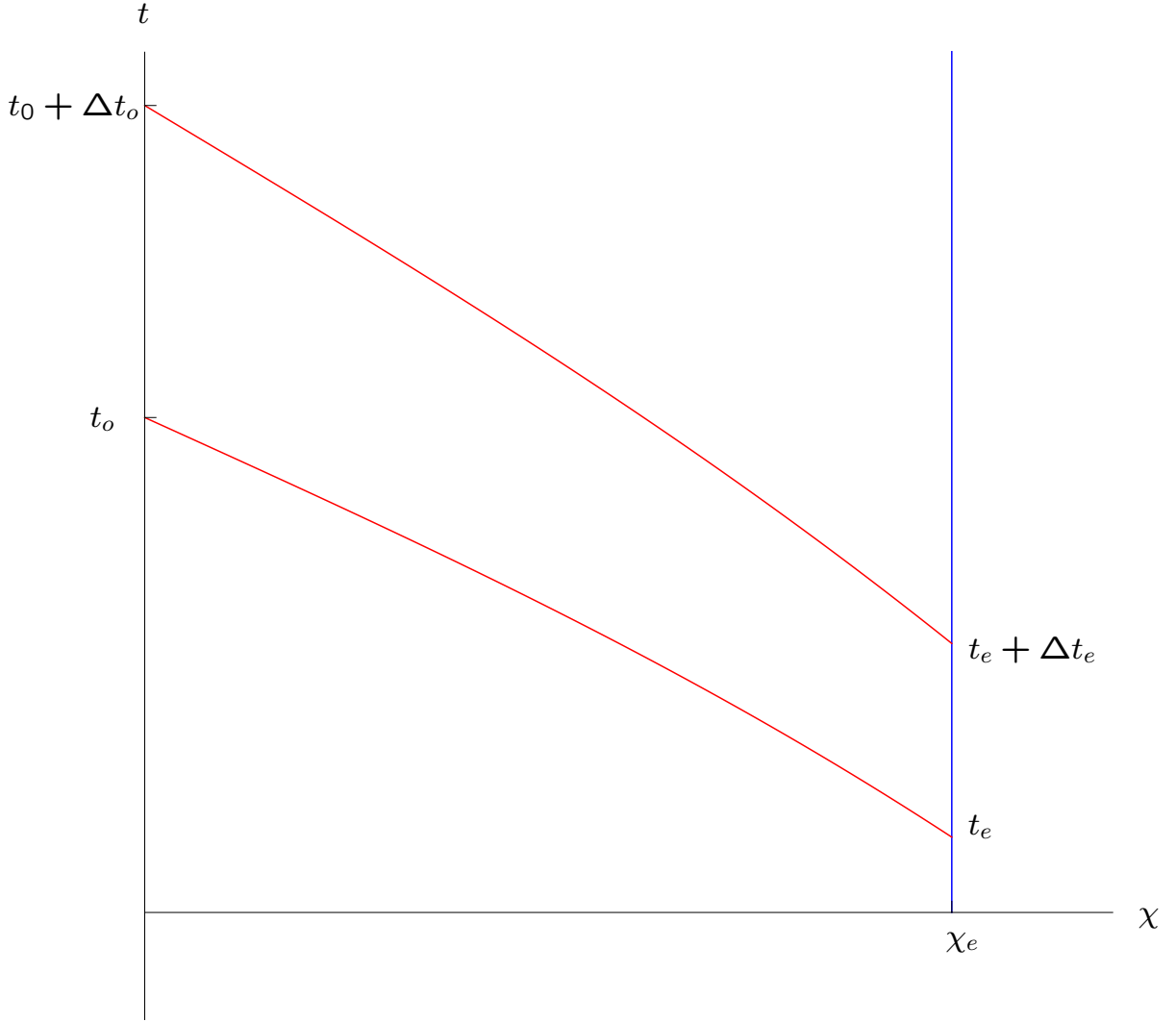


Figure 10: Redshift for standard observers in a Robertson-Walker spacetime

With a Taylor expansion

$$a(t_e) = a(t_o) + \frac{da}{dt}(t_o)(t_e - t_o) + \frac{1}{2} \frac{d^2a}{dt^2}(t_o)(t_e - t_o)^2 + \dots \quad (72)$$

we find

$$\begin{aligned} z &= \frac{a(t_o) - a(t_e)}{a(t_e)} = \frac{-\frac{da}{dt}(t_o)(t_e - t_o) - \frac{1}{2} \frac{d^2a}{dt^2}(t_o)(t_e - t_o)^2 + \dots}{a(t_o) + \frac{da}{dt}(t_o)(t_e - t_o) + \dots} = \\ &= \frac{\frac{da}{dt}(t_o)(t_o - t_e) - \frac{1}{2} \frac{d^2a}{dt^2}(t_o)(t_o - t_e)^2 + \dots}{a(t_o) \left(1 - \frac{da}{dt}(t_o) \frac{(t_o - t_e)}{a(t_o)} + \dots \right)} = \end{aligned}$$

$$\begin{aligned}
&= \left(\frac{da}{dt}(t_o)(t_o - t_e) - \frac{1}{2} \frac{d^2a}{dt^2}(t_o)(t_o - t_e)^2 + \dots \right) \frac{1}{a(t_o)} \left(1 + \frac{da}{dt}(t_o) \frac{(t_o - t_e)}{a(t_o)} + \dots \right) = \\
&= \frac{1}{a(t_o)} \frac{da}{dt}(t_o)(t_o - t_e) + \left(\frac{1}{a(t_o)^2} \left(\frac{da}{dt}(t_o) \right)^2 - \frac{1}{2a(t_o)} \frac{d^2a}{dt^2}(t_o) \right) (t_o - t_e)^2 + \dots \quad (73)
\end{aligned}$$

It is common to define the *Hubble constant*

$$H(t_o) := \frac{1}{a(t_o)} \frac{da}{dt}(t_o) \quad (74)$$

and the *deceleration parameter*

$$q(t_o) := \frac{-a(t_o)}{\left(\frac{da}{dt}(t_o) \right)^2} \frac{d^2a}{dt^2}(t_o). \quad (75)$$

Then the expression for the redshift becomes

$$z = H(t_o)(t_o - t_e) + H(t_o)^2 \left(1 + \frac{q(t_o)}{2} \right) (t_o - t_e)^2 + \dots \quad (76)$$

The travel time $t_o - t_e$ can be viewed as a measure for the distance (if multiplied with c , for dimensional reasons). We will discuss other distance measures below in this section and we will see that all of them coincide up to first order in $t_o - t_e$. So we may say that the relation between distance and redshift is unambiguously determined by the Hubble constant to within a linear approximation. Note that the Hubble “constant” and the deceleration parameter depend on t_o . We will discuss later in detail that we believe to live in a universe where now (at time t_o) the Hubble constant is positive and the deceleration parameter is negative, i.e., in a universe that is expanding and where the expansion rate is even increasing.

- (ii) **Horizons:** There are two types of horizons in a Robertson-Walker universe, known as “particle horizons” and “event horizons”. Here the word “particle” is used as synonymous with “standard observer”. These notions are defined in the following way.

Fix an event p_o . Then the *particle horizon* of p_o separates particles that can be seen at p_o from those that cannot.

Fix a particle (i.e., a standard observer) P_o . Then the *event horizon* of P_o separates events that can be seen by P_o from those that cannot.

Here p_o is a point in the spacetime while P_o is a worldline. Always keep in mind that events have particle horizons whereas particles have event horizons.

It is now our goal to give a mathematical criterion for the existence or non-existence of horizons. To that end we perform a coordinate transformation where only the time coordinate is transformed, $(t, \chi, \vartheta, \varphi) \mapsto (\eta, \chi, \vartheta, \varphi)$. The new time coordinate η is dimensionless, and it is defined by

$$\eta = \int_{t_o}^t \frac{c d\tilde{t}}{a(\tilde{t})} \quad (77)$$

where t_o is a constant that can be chosen at will, except for the restriction $t_i < t_o < t_f$. Then t_o corresponds to $\eta_o = 0$ while t_i and t_f correspond to

$$\eta_i = \int_{t_o}^{t_i} \frac{c d\tilde{t}}{a(\tilde{t})} = - \int_{t_i}^{t_o} \frac{c d\tilde{t}}{a(\tilde{t})} < 0 \quad (78)$$

and

$$\eta_f = \int_{t_o}^{t_f} \frac{c d\tilde{t}}{a(\tilde{t})} > 0, \quad (79)$$

respectively. If we express the scale factor as a function of η ,

$$a(t) = \hat{a}(\eta), \quad (80)$$

the metric reads

$$g = \hat{a}(\eta)^2 \left(-d\eta^2 + d\chi^2 + \xi(\chi)^2 \{ d\vartheta^2 + \sin^2\vartheta d\varphi^2 \} \right). \quad (81)$$

Note that inside the big bracket all quantities are dimensionless. Multiplying a metric with a positive function is called a *conformal transformation*. We see that, up to a dimensional factor, η is proper time along the worldlines of the standard observers with respect to the conformally transformed metric $\hat{a}(\eta)^{-2}g$; therefore, η is called the *conformal time*. As the coefficients of the metric $\hat{a}(\eta)^{-2}g$ are independent of η , this metric is static. The given representation thus shows that every Robertson-Walker metric is *conformally static*, i.e., conformal to a static metric.

If we use the conformal time coordinate η instead of t , equation (66) for lightlike radial geodesics reads

$$\chi = \pm \eta, \quad (82)$$

i.e., in a (χ, η) -diagram the radial light rays are represented by lines under 45 degrees.

It is now obvious that the following existence criteria for horizons are true, see Figure 11 below.

Particle horizons exist if and only if $|\eta_i| < \chi_{\max}$.

Event horizons exist if and only if $\eta_f < \chi_{\max}$.

In particular, horizons do not exist if $\eta_i = -\infty$ and $\eta_f = \infty$, i.e., if the function $\eta \mapsto \hat{a}(\eta)$ is defined on all of \mathbb{R} . Note that the proper time coordinate t is relevant for the question of whether there is a big bang and/or a big crunch (or a big rip); the conformal time coordinate η is relevant for the question of whether there are horizons.

Figure 11 shows an event horizon and a particle horizon for a case where η_i , η_f and χ_{\max} are finite.

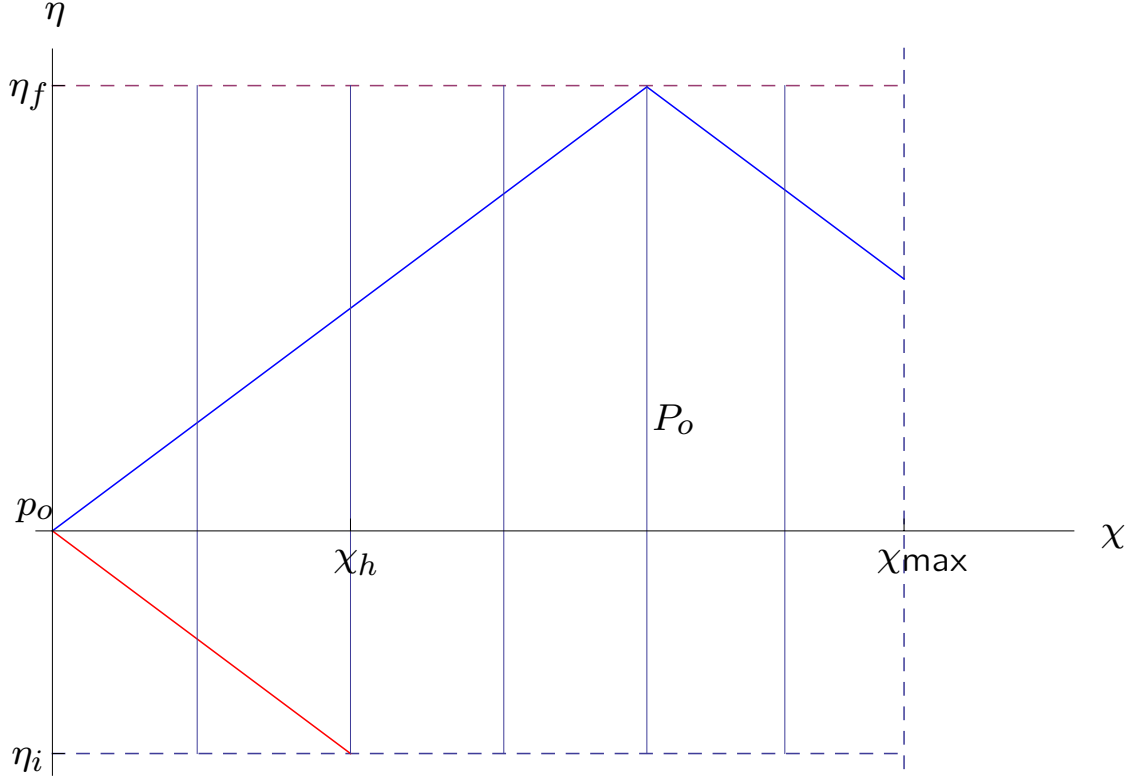


Figure 11: Particle horizon of an event p_o and event horizon of a particle P_o in a Robertson-Walker spacetime. The past light-cone of the event p_o is shown in red, so the particle horizon of p_o is situated at the radius coordinate χ_h . The event horizon of the particle P_o is shown in blue.

- (iii) Distance measures: There are various ways of assigning a distance to a pair of standard observers. We will discuss several of them, interpreting one of the standard observers as the emitter of light and the other as the observer. As the scale factor depends on time, each of the distance measures has to be viewed as a function of the observation time t_o . We choose the observer as the origin of the spatial coordinate system unless otherwise stated. Astronomers prefer to use the redshift z as an independent variable because it is directly measurable (if spectral lines can be identified). Following this practice, we write each of the distance measures as a power series in terms of z . Such series are known as “Kristian-Sachs series”. We will see that all distance measures coincide to within linear order.

– Distance by travel time of light

If a light ray starts at time t_e at the emitter and arrives at time t_o at the observer, we can use the expression

$$D_T = c(t_o - t_e) \quad (83)$$

as a measure for the distance between emitter and observer. D_T is not a directly measurable quantity because t_e is not known. However, in some cases t_e and thus D_T can be estimated.

With the series expansion (76) for the redshift we can express D_T as a power series in terms of z . For that purpose, we write

$$t_o - t_e = \alpha z + \beta z^2 + \dots \quad (84)$$

and insert this expression into (76),

$$\begin{aligned} z &= H(t_o) \left(\alpha z + \beta z^2 \right) + H(t_o)^2 \left(1 + \frac{q(t_o)}{2} \right) \alpha^2 z^2 + \dots = \\ &= H(t_o) \alpha z + \left\{ H(t_o) \beta + H(t_o)^2 \left(1 + \frac{q(t_o)}{2} \right) \alpha^2 \right\} z^2 + \dots \end{aligned} \quad (85)$$

Comparison of coefficients yields

$$1 = H(t_o) \alpha, \quad 0 = H(t_o) \left\{ \beta + H(t_o) \left(1 + \frac{q(t_o)}{2} \right) \alpha^2 \right\} \quad (86)$$

hence

$$\alpha = \frac{1}{H(t_o)}, \quad \beta = -\frac{1}{H(t_o)} \left(1 + \frac{q(t_o)}{2} \right). \quad (87)$$

Upon reinserting these expressions into (84), we get $D_T = c(t_o - t_e)$ as a power series in terms of z ,

$$D_T = \frac{c z}{H(t_o)} - \left(1 + \frac{q(t_o)}{2} \right) \frac{c z^2}{H(t_o)} + \dots \quad (88)$$

– Proper distance

From a mathematical point of view, the most natural way of measuring the distance between two standard observers (emitter and observer) at time t_o is the proper length of the radial line that connects them in the hypersurface $t = t_o$. From the form of the metric

$$g = -c^2 dt^2 + a(t)^2 \left(d\chi^2 + \xi(\chi)^2 (d\vartheta^2 + \sin^2 \vartheta d\varphi^2) \right), \quad (89)$$

we read that along a radial line ($t = t_o$, $\vartheta = \text{constant}$, $\varphi = \text{constant}$) proper length ℓ is given by

$$d\ell^2 = a(t_o)^2 d\chi^2, \quad (90)$$

so the proper distance between an emitter at $\chi_e = \chi$ and an observer at $\chi_o = 0$ is

$$D_p = \int_0^\chi a(t_o) d\chi = a(t_o) \chi. \quad (91)$$

From this expression we find a new version of the Hubble law,

$$\frac{dD_p}{dt}(t_o) = \frac{da}{dt}(t_o) \chi = \frac{da}{dt}(t_o) \frac{D_p}{a(t_o)} = H(t_o) D_p. \quad (92)$$

We may interpret this formula as saying that the radial velocity of a distant source is proportional to its distance, with the Hubble constant as the factor of proportionality. This formula is exact, i.e., not only valid to within a linear approximation. In this sense, it is more universal than the linear relation between travel time and redshift which is true only if terms of higher-order in the travel time are neglected, see (76).

On the other hand, the “radial velocity” and the “distance” in this formula are purely theoretical quantities that are not related to observations: The proper distance D_p is based on connecting two events in a hypersurface $t = t_o$, i.e., at equal times; no signal can realise this connecting line. Moreover, the “radial velocity” dD_p/dt does *not* give the velocity of the emitter relative to its neighbourhood but rather the change of a mathematically defined distance between emitter and observer. Therefore, it is no contradiction to the rules of relativity that this “radial velocity” may very well be bigger than c , see Worksheet 3 for an example.

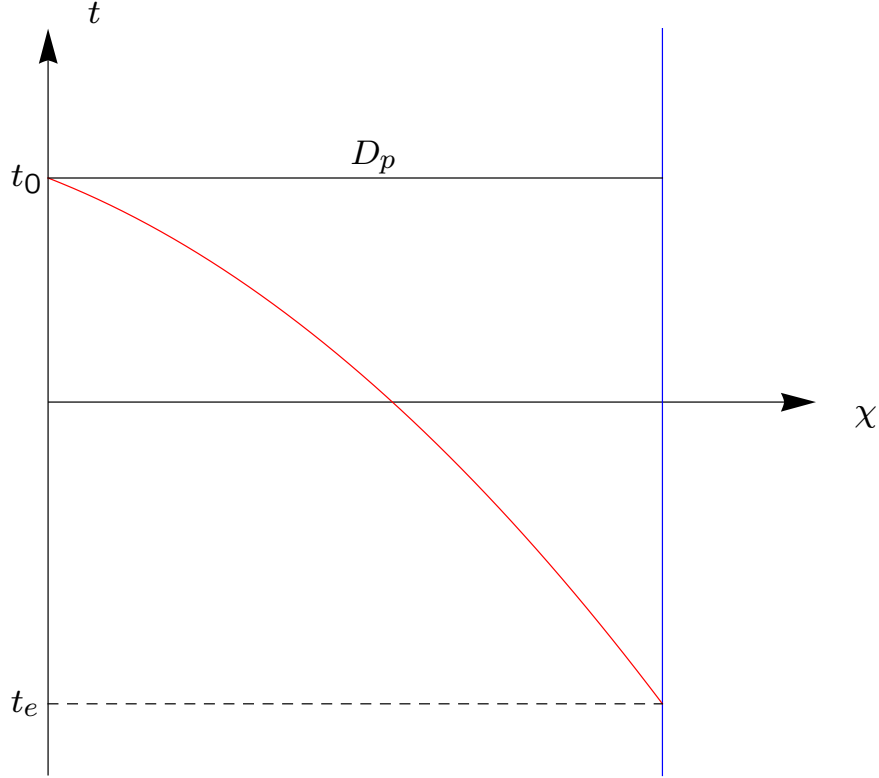


Figure 12: Proper distance between two standard observers in a Robertson-Walker spacetime

In view of the fact that the Hubble constant relates a distance to a velocity it is usually given in units of (km/s)/Mpc. Of course, this is the same as an inverse time.

As in the case of the distance by travel time of light, we may write D_p as a series in terms of the redshift z . In this case we need the Taylor expansion of χ using eq. (66) for a future-oriented radial light ray,

$$\begin{aligned}\chi &= \int_{t_e}^{t_o} \frac{c dt}{a(t)} = \int_{t_e}^{t_o} \frac{c dt}{a(t_o) + \frac{da}{dt}(t_o)(t - t_o) + \dots} = \\ &= \int_{t_e}^{t_o} \frac{c dt}{a(t_o) \left(1 + H(t_o)(t - t_o) + \dots \right)} =\end{aligned}$$

$$\begin{aligned}
&= \frac{c}{a(t_o)} \int_{t_e}^{t_o} \left(1 - H(t_o)(t - t_o) + \dots \right) dt = \\
&= \frac{c}{a(t_o)} \left\{ t_o - t_e - H(t_o) \left(\frac{t_o^2}{2} - \frac{t_e^2}{2} - t_o(t_o - t_e) \right) + \dots \right\} = \\
&= \frac{c}{a(t_o)} (t_o - t_e) + \frac{c}{2a(t_o)} H(t_o)(t_o - t_e)^2 + \dots
\end{aligned} \tag{93}$$

Inserting the Taylor expansion (88) of $D_T = c(t_o - t_e)$ in terms of z yields

$$\begin{aligned}
D_p = a(t_o)\chi &= \frac{cz}{H(t_o)} - \left(1 + \frac{q(t_o)}{2} \right) \frac{cz^2}{H(t_o)} + \frac{cz^2}{2H(t_o)} + \dots = \\
&= \frac{cz}{H(t_o)} - \frac{c}{2H(t_o)} (1 + q(t_o)) z^2 + \dots
\end{aligned} \tag{94}$$

– Area distance (=angular diameter distance)

In Newtonian physics, which is based on the assumption that Euclidean geometry holds in our 3-space, the apparent size of an object is inverse proportional to the square of its distance. If we have “standard rulers” (i.e., objects whose true size we know) at our disposal, we can determine their distance directly from measuring their apparent size in the sky. In a curved geometry, we can *define* a distance measure in such a way that this relation still holds. This can be done either by comparing the true cross-sectional area of the object to the solid angle it suspends in the sky, or by comparing the true length of a particular diameter of the object to the angle this diameter suspends in the sky. The first distance measure is known as the “area distance” and the second as the “angular diameter distance”. In a Robertson-Walker universe the two notions coincide because of the isotropy. Moreover, isotropy implies that it suffices to consider the area of spheres about the observer. So we choose the observer as the spatial origin, as before, and we consider the past light-cone of an observation event at time t_o , see Figure 13. For any earlier time t_e , the intersection of this light-cone with the hypersurface $t = t_e$ is a sphere of coordinate radius

$$\chi = \int_{t_e}^{t_o} \frac{c dt}{a(t)}. \tag{95}$$

From the metric we read that the area of this sphere is $4\pi a(t_e)^2 \xi(\chi)^2$. The area distance D_A is *defined* by equating this expression to the Euclidean expression for the area of a sphere,

$$4\pi a(t_e)^2 \xi(\chi)^2 = 4\pi D_A^2, \tag{96}$$

hence

$$D_A = a(t_e) \xi(\chi). \tag{97}$$

As $\xi(\chi) = \sin \chi$, $\xi(\chi) = \chi$ or $\xi(\chi) = \sinh \chi$, we have in any case $\xi(\chi) = \chi + O(\chi^3)$, so we may write

$$D_A = a(t_e) \left(\chi + O(\chi^3) \right). \tag{98}$$

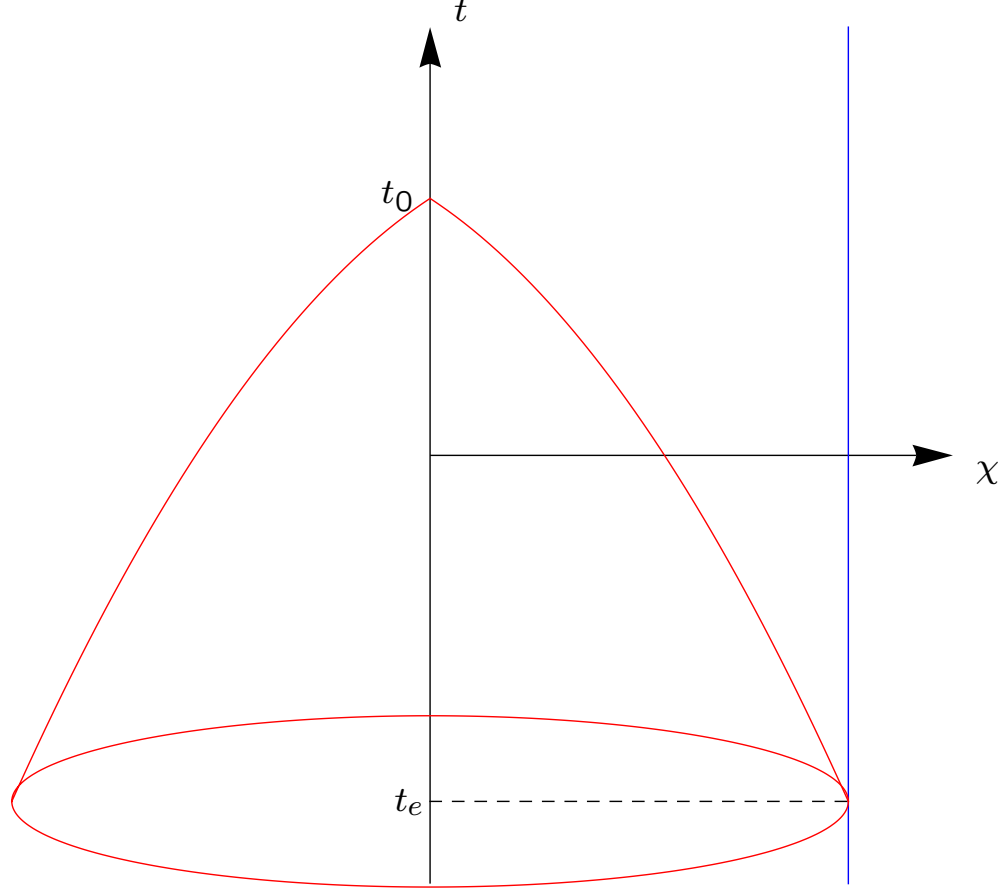


Figure 13: Area distance between two standard observers in a Robertson-Walker spacetime

With the Taylor series (93) for χ this can be rewritten as

$$\begin{aligned}
 D_A &= \frac{a(t_e)}{a(t_o)} a(t_o) \left(\frac{c}{a(t_o)} (t_o - t_e) + \frac{c H(t_o)}{2a(t_o)} (t_o - t_e)^2 + \dots \right) \\
 &= \frac{1}{(1+z)} \left(D_T + \frac{H(t_o)}{2c} D_T^2 + \dots \right)
 \end{aligned} \tag{99}$$

Inserting the Taylor series (88) for D_T yields

$$\begin{aligned}
 D_A &= (1 - z + \dots) \left(\left\{ \frac{cz}{H(t_o)} - \left(1 + \frac{q(t_o)}{2}\right) \frac{cz^2}{H(t_o)} + \dots \right\} + \frac{H(t_o)}{2} \frac{cz^2}{H(t_o)^2} + \dots \right) \\
 &= \frac{cz}{H(t_o)} - (3 + q(t_o)) \frac{cz^2}{2H(t_o)} + \dots
 \end{aligned} \tag{100}$$

If we had standard rulers available distributed over the universe, we could measure D_A and z for each of them and then determine $H(t_o)$ and $q(t_o)$ from this formula. Actually, we have better “standard candles” than “standard rulers”, i.e., it is more promising to consider the luminosity of a light source rather than its size, see the next item.

– Luminosity distance

In Newtonian physics, not only the apparent size but also the apparent luminosity of a light source falls off with the square of the distance. So if we have “standard candles” (i.e., objects whose true luminosity we know) at our disposal, we can determine their distance directly from measuring their apparent luminosity. The apparent luminosity is given by the energy flux arriving at the observer.

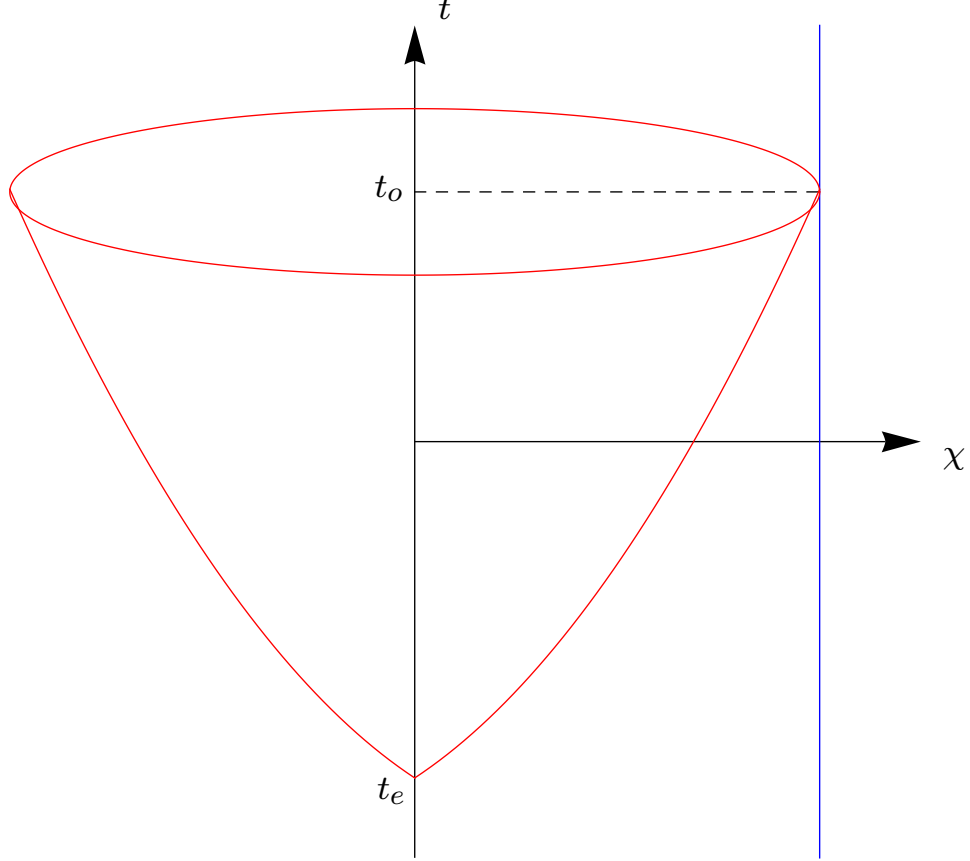


Figure 14: (Corrected) luminosity distance between two standard observers in a Robertson-Walker spacetime

In analogy to the area distance, which was defined in a way that it is related to the true size of a light source by the same formula as in the Newtonian (i.e., Euclidean) case, we can define a luminosity distance in a way that it is related to the true luminosity by the Newtonian formula. In a Robertson-Walker universe, we may again take advantage of the isotropy and consider the future light-cone of an emission event at time t_e . In this case it is convenient to place the emitter in the spatial origin, $\chi_e = 0$, and to have the observer at a radius coordinate $\chi_o = \chi$, see Figure 14. Then for any observation time $t_o > t_e$, the intersection of the considered future light-cone with the hypersurface $t = t_o$ has area $4\pi a(t_o)^2 \xi(\chi)^2$. We assume that the emitter sends photons isotropically into all spatial directions. If the apparent luminosity were measured in terms of a number flux of photons, the desired distance measure would be given by the equation

$$4\pi a(t_o)^2 \xi(\chi)^2 = 4\pi \tilde{D}_L^2. \quad (101)$$

One calls \tilde{D}_L the *corrected luminosity distance*. Actually, the apparent luminosity is not given by the number flux of photons but rather by the energy flux. The latter differs from the first by a redshift factor, because the energy of a photon undergoes a redshift on its way from the emitter to the observer. Therefore, one defines the (uncorrected) *luminosity distance* D_L by the equation

$$D_L = (1 + z) \tilde{D}_L = a(t_o)(1 + z) \xi(\chi). \quad (102)$$

This is related to the energy flux F at the observer by the formula

$$F = \frac{L}{4\pi D_L^2} \quad (103)$$

where L is the true luminosity of the source. (One usually considers the luminosity integrated over all frequencies which is called the *bolometric luminosity*.) Astronomers use a logarithmic scale, as all human senses respond logarithmically to a physical stimulus (“Weber-Fechner law”) and define the (apparent) *magnitude* m of a light source such that

$$m = -2.5 \log_{10}(L) + 2.5 \log_{10}(D_L^2) + m_0 \quad (104)$$

where m_0 is a constant. Note that the luminosity distance can be rewritten as

$$D_L = \frac{a(t_o)}{a(t_e)} a(t_e) (1 + z) \xi(\chi) = (1 + z)^2 a(t_e) \xi(\chi). \quad (105)$$

Comparison with eq. (97) for the area distance demonstrates that

$$D_L = (1 + z)^2 D_A. \quad (106)$$

This relation between luminosity distance and area distance was proven here for a Robertson-Walker universe. Actually, it is true in *any* spacetime, but the general proof is rather involved and will not be given here. It is based on the so-called *reciprocity theorem* for light bundles which was proven by Etherington in 1933.

With the relation between luminosity distance and area distance at hand, we can now easily write the luminosity distance as a power series in terms of z ,

$$\begin{aligned} D_L &= (1 + 2z + z^2) \left(\frac{cz}{H(t_o)} - (3 + q(t_o)) \frac{cz^2}{2H(t_o)} + \dots \right) = \\ &= \frac{cz}{H(t_o)} + (1 - q(t_o)) \frac{cz^2}{2H(t_o)} + \dots \end{aligned} \quad (107)$$

If we have standard candles distributed in the universe, we can measure their luminosity distance and their redshift and determine $H(t_o)$ and $q(t_o)$ from the last equation. The best standard candles we have to date are supernovae of type Ia. In the next chapter we will discuss in detail how they have been used to determine $q(t_o)$ in the late 1990s. The surprising result was that $q(t_o) < 0$, i.e., that the expansion of our universe is accelerating. The present value of the Hubble constant

is $H(t_o) = (67.8 \pm 0.77)(\text{km/s})/\text{Mpc}$. This, however, was not determined from the relation between luminosity distance and redshift, which yields a considerably lower accuracy, but rather from the cosmic background radiation, see below.

For each of the distance measures we have discussed the series expansion in terms of z is true in *any* Robertson-Walker universe. The results are purely kinematical in the sense that Einstein's field equation has not been used. Also note that our formulas, which include terms up to the second order, are independent of k , i.e., they hold for universes of $k = 1$, $k = 0$ and $k = -1$. The third-order terms, however, do depend on k , at least for D_A and D_L , because then the $O(\chi^3)$ term in $\xi(\chi) = \chi + O(\chi^3)$ has to be taken into account. To date the third-order terms are beyond the reach of observations.

3.2 Friedmann-Lemaître solutions

In this section we will study those Robertson-Walker spacetimes that satisfy Einstein's field equation with a perfect fluid source. We will see that *any* Robertson-Walker spacetime can be viewed as such a solution if we do not impose conditions on the energy density or the pressure. However, in view of applications to cosmology it is of particular interest to study solutions where certain properties of the energy density or the pressure have been specified. This is what one calls the Friedmann-Lemaître solutions.

It is convenient to use Robertson-Walker spacetimes in the coordinate representation

$$g = -c^2 dt^2 + a(t)^2 \left(\frac{d\xi^2}{1 - k \xi^2} + \xi^2 (d\vartheta^2 + \sin^2 \vartheta d\varphi^2) \right). \quad (108)$$

Writing a prime for the derivative of the scale factor with respect to the time coordinate t , one finds that the components of the Ricci tensor are

$$R_{tt} = -\frac{3}{a(t)} a''(t), \quad (109)$$

$$R_{\xi\xi} = \frac{1}{(1 - k \xi^2)} \left(2k + \frac{2}{c^2} a'(t)^2 + \frac{a(t)}{c^2} a''(t) \right), \quad (110)$$

$$R_{\varphi\varphi} = \sin^2 \vartheta R_{\vartheta\vartheta} = \xi^2 (1 - k \xi^2) R_{\xi\xi}, \quad (111)$$

and $R_{\mu\nu} = 0$ for $\mu \neq \nu$. As a consequence, the Ricci scalar reads

$$R = \frac{6}{c^2 a(t)^2} \left(c^2 k + a'(t)^2 + a(t) a''(t) \right). \quad (112)$$

We want to solve Einstein's field equation with an energy-momentum tensor of a perfect fluid,

$$T_{\mu\nu} = \left(\varepsilon + p \right) \frac{U_\mu U_\nu}{c^2} + p g_{\mu\nu} \quad (113)$$

where the fluid is supposed to be at rest with respect to the standard observers. The latter condition requires, in combination with the normalisation condition $U^\mu U_\mu = -c^2$, that

$$U^\mu = \delta_t^\mu, \quad U_\nu = g_{\nu t} = -c^2 \delta_\nu^t. \quad (114)$$

We interpret U^μ as the 4-velocity of the mean flow of galaxies.

For the components of the energy-momentum tensor we find

$$T_{tt} = \left(\varepsilon + p \right) \frac{(g_{tt})^2}{c^2} + p g_{tt} = \varepsilon c^2, \quad (115)$$

$$T_{\xi\xi} = p g_{\xi\xi} = \frac{p a(t)^2}{1 - k \xi^2}, \quad (116)$$

$$T_{\varphi\varphi} = \sin^2 \vartheta T_{\vartheta\vartheta} = \xi^2 (1 - k \xi^2) \sin^2 \vartheta T_{\xi\xi}. \quad (117)$$

For $\mu \neq \nu$ we have $T_{\mu\nu} = 0$.

Einstein's field equation

$$R_{\mu\nu} - \frac{R}{2} g_{\mu\nu} + \Lambda g_{\mu\nu} = \kappa T_{\mu\nu} \quad (118)$$

gives two independent component equations. The tt component yields

$$\begin{aligned} -\frac{3}{a(t)} a''(t) + \frac{3c^2}{c^2 a(t)^2} \left(c^2 k + a'(t)^2 + a(t) a''(t) \right) - \Lambda c^2 &= \kappa \varepsilon c^2, \\ \frac{3k}{a(t)^2} + \frac{3}{c^2} \frac{a'(t)^2}{a(t)^2} - \Lambda &= \kappa \varepsilon, \end{aligned} \quad (119)$$

and the $\xi\xi$ component yields

$$\begin{aligned} \frac{1}{(1 - k \xi^2)} \left(2k + \frac{2}{c^2} a'(t)^2 + \frac{a(t)}{c^2} a''(t) \right) - \frac{3a(t)^2}{c^2 a(t)^2 (1 - k \xi^2)} \left(c^2 k + a'(t)^2 + a(t) a''(t) \right) \\ + \frac{\Lambda a(t)^2}{1 - k \xi^2} = \frac{\kappa p a(t)^2}{1 - k \xi^2}, \\ -\frac{k}{a(t)^2} - \frac{1}{c^2} \frac{a'(t)^2}{a(t)^2} - \frac{2}{c^2} \frac{a''(t)}{a(t)} + \Lambda = \kappa p. \end{aligned} \quad (120)$$

The $\varphi\varphi$ and $\vartheta\vartheta$ components give again eq. (120) and the $\mu\nu$ components with $\mu \neq \nu$ just give the identity $0 = 0$.

(119) and (120) are known as the (generalised) *Friedmann equations*. They were found by Alexander Friedmann in 1921/22 for the dust case ($p = 0$) and generalised by Georges Lemaître in 1927 to the case with $p \neq 0$. As the left-hand sides are functions of t only, these equations require that the energy density ε and the pressure p be also functions of t only.

Moreover, we read from (119) and (120) that $a(t)$, k and Λ can be chosen arbitrarily and then $\varepsilon(t)$ and $p(t)$ are uniquely determined. In this sense, *any* Robertson-Walker spacetime solves Einstein's field equation with a perfect fluid source that is at rest with respect to the standard observers. However, it is *not* guaranteed that pressure and density are non-negative and related by an equation of state, i.e., by a relation of the form $F(\varepsilon, p) = 0$ that does not depend explicitly on time. So what one has to discuss is the question of which Robertson-Walker spacetimes are *physically reasonable* perfect fluid solutions.

(a) Vacuum solutions

Although our real universe is certainly not a vacuum spacetime, vacuum solutions to the Friedmann equations are of some interest in cosmology as limiting cases. For vacuum, $\mu = 0$ and $p = 0$, the Friedmann equations reduce to

$$c^2 k + a'(t)^2 - \frac{\Lambda}{3} c^2 a(t)^2 = 0, \quad (121)$$

$$-c^2 k - a'(t)^2 - 2 a(t) a''(t) + \Lambda c^2 a(t)^2 = 0. \quad (122)$$

We will first show that in the vacuum case the second equation is essentially redundant.

Claim: If $a'(t) \neq 0$, (122) is a consequence of (121).

Proof: Differentiation of (121) yields

$$2 a'(t) a''(t) - \frac{\Lambda}{3} c^2 2 a(t) a'(t) = 0.$$

As we assume $a'(t) \neq 0$, we may multiply with $a(t)/a'(t)$. This yields

$$2 a(t) a''(t) = \frac{\Lambda}{3} c^2 2 a(t)^2.$$

Using again (121), this can be rewritten as

$$2 a(t) a''(t) + a'(t)^2 = \frac{\Lambda}{3} c^2 2 a(t)^2 - c^2 k + \frac{\Lambda}{3} c^2 a(t)^2$$

which is just equation (122).

If $a'(t)$ has isolated zeros, (122) is a consequence of (121) as well, because of continuity. So the only case which has to be considered separately is the case that the scale factor is constant on the considered interval, $a(t) = a_0$. Then the sum of (121) and (122) requires $\Lambda = 0$; thereupon, (121) requires $k = 0$. So in all cases except if $\Lambda = 0$, $k = 0$ and $a(t) = a_0$ is it sufficient to consider (121). We now consider all possible cases for Λ and for k one by one.

(i) $\Lambda = 0$

• $k = 1$

In this case the Friedmann equation (121) reads

$$c^2 + a'(t)^2 = 0$$

which cannot be satisfied by any real function a .

- $k = 0$

We know that $a(t) = a_0$ is a solution. We have to check if it is the only solution or if (121) admits an additional solution. As a matter of fact, in the case at hand (121) reduces to

$$a'(t)^2 = 0,$$

so $a(t) = a_0$ is indeed the only solution. In this case the metric reads

$$g = -c^2 dt^2 + a_0^2 \left(d\chi^2 + \chi^2 (d\vartheta^2 + \sin^2 \vartheta d\varphi^2) \right). \quad (123)$$

A coordinate transformation $(t, \chi, \vartheta, \varphi) \mapsto (t, r = a_0 \chi, \vartheta, \varphi)$ shows that this is just the Minkowski metric,

$$g = -c^2 dt^2 + dr^2 + r^2 (d\vartheta^2 + \sin^2 \vartheta d\varphi^2).$$

The time slices are the usual flat Euclidean spaces of an inertial system, represented in spherical polar coordinates.

- $k = -1$

Now (121) requires solving the differential equation

$$-c^2 + a'(t)^2 = 0$$

which yields

$$\frac{da}{dt} = \pm c, \quad a(t) = \pm c(t - t_0).$$

As we are free to shift the zero point on the time axis, we may choose $t_0 = 0$. Moreover, we only consider the plus sign because the minus sign gives the same spacetime just with the time reversed. The scale factor $a(t) = ct$ is then defined on the interval $]t_i = 0, t_f = \infty[$. The metric reads

$$g = -c^2 dt^2 + c^2 t^2 \left(d\chi^2 + \sinh^2 \chi (d\vartheta^2 + \sin^2 \vartheta d\varphi^2) \right) \quad (124)$$

which is *Milne's universe* that was already mentioned. Milne's universe is just part of Minkowski spacetime, with a slicing into hyperboloids, see Figure 8. The transformation to Minkowski spacetimes in spherical polars $(\tilde{t}, \tilde{r}, \vartheta, \varphi)$ is given by

$$\tilde{t} = t \cosh \chi, \quad \tilde{r} = ct \sinh \chi. \quad (125)$$

(ii) $\Lambda > 0$

- $k = 1$

In this case the Friedmann vacuum equation (121) reads

$$c^2 + a'(t)^2 - \frac{\Lambda}{3} c^2 a(t)^2 = 0,$$

$$\frac{da}{dt} = \pm c \sqrt{\frac{\Lambda}{3} a^2 - 1}.$$

As Λ is positive, we may substitute

$$\sqrt{\frac{\Lambda}{3}} a = \cosh u, \quad \sqrt{\frac{\Lambda}{3}} da = \sinh u du,$$

hence

$$\sqrt{\frac{3}{\Lambda}} \frac{\sinh u du}{\sqrt{\cosh^2 u - 1}} = \pm c dt.$$

This results in

$$u = \pm \sqrt{\frac{\Lambda}{3}} c (t - t_0),$$

$$a(t) = \sqrt{\frac{3}{\Lambda}} \cosh \left(\sqrt{\frac{\Lambda}{3}} c (t - t_0) \right). \quad (126)$$

Without loss of generality we choose $t_0 = 0$. The scale factor is defined for all real values of t . It decreases from $t = -\infty$ to a minimum value at $t = 0$ and then increases again to $+\infty$. The metric reads

$$g = -c^2 dt^2 + \frac{3}{\Lambda} \cosh^2 \left(\sqrt{\frac{\Lambda}{3}} c t \right) (d\chi^2 + \sin^2 \chi (d\vartheta^2 + \sin^2 \vartheta d\varphi^2)). \quad (127)$$

This spacetime is known as the *deSitter universe*. It was found by Dutch astronomer Willem deSitter in 1917. The deSitter universe can be isometrically embedded as the hyperboloid

$$X^2 + Y^2 + Z^2 + W^2 - V^2 = \frac{3}{\Lambda} \quad (128)$$

into 5-dimensional Minkowski space,

$$g^{(5)} = dX^2 + dY^2 + dZ^2 + dW^2 - dV^2. \quad (129)$$

In terms of our coordinates $(t, \chi, \vartheta, \varphi)$, the embedding is given by the map

$$X = \sqrt{\frac{3}{\Lambda}} \cosh \left(\sqrt{\frac{\Lambda}{3}} c t \right) \sin \chi \cos \varphi \sin \vartheta, \quad (130)$$

$$Y = \sqrt{\frac{3}{\Lambda}} \cosh \left(\sqrt{\frac{\Lambda}{3}} c t \right) \sin \chi \sin \varphi \sin \vartheta, \quad (131)$$

$$Z = \sqrt{\frac{3}{\Lambda}} \cosh \left(\sqrt{\frac{\Lambda}{3}} c t \right) \sin \chi \cos \vartheta, \quad (132)$$

$$W = \sqrt{\frac{3}{\Lambda}} \cosh \left(\sqrt{\frac{\Lambda}{3}} c t \right) \cos \chi, \quad (133)$$

$$V = \sqrt{\frac{3}{\Lambda}} \sinh \left(\sqrt{\frac{\Lambda}{3}} c t \right). \quad (134)$$

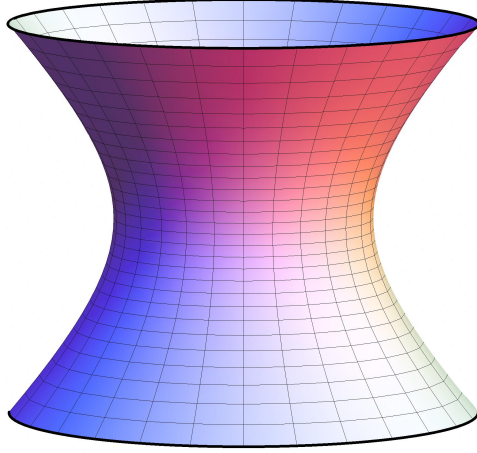


Figure 15: The deSitter universe embedded as a hyperboloid in 5-dimensional Minkowski spacetime, with a slicing into 3-spheres

Our representation gives the deSitter universe with a slicing into hyperspaces $t = \text{constant}$ that are 3-spheres S^3 , so the topology of the spacetime is $S^3 \times \mathbb{R}$. In Figure 15 two spatial dimensions are suppressed, so the 3-spheres are represented by circles which are given by intersecting the hyperboloid with horizontal planes.

We will briefly check if there are horizons in the deSitter universe with its sclicing into 3-spheres. To that end we have to consider the conformal time η which is defined by

$$d\eta = \frac{c dt}{a(t)} = c \sqrt{\frac{\Lambda}{3}} \frac{dt}{\cosh\left(\sqrt{\frac{\Lambda}{3}} ct\right)}. \quad (135)$$

Integration yields

$$\eta - \eta_0 = \arccos\left(\tanh\left(\sqrt{\frac{\Lambda}{3}} ct\right)\right). \quad (136)$$

The conformal time η is then related to t by

$$\cos(\eta - \eta_0) = \tanh\left(\sqrt{\frac{\Lambda}{3}} ct\right). \quad (137)$$

If t runs over its domain from $-\infty$ to ∞ , the dimensionless conformal time parameter η runs from $\eta_i = \eta_0 - \pi$ to $\eta_f = \eta_0$. For every event in the spacetime, $\eta - \eta_i$ is smaller than $\chi_{\max} = \pi$, so there are particle horizons, recall Figure 11. The part of the 3-sphere that is visible becomes bigger and bigger for $t \rightarrow \infty$ and the antipodal point is the only point that comes never into view. Similarly, $\eta_f - \eta$ is smaller than χ_{\max} , so there are event horizons.

The deSitter universe is of great relevance for cosmology. We will see later that it plays an important role for the ideas of inflation and of dark energy.

- $k = 0$

Now (121) requires solving the differential equation

$$a'(t)^2 - \frac{\Lambda}{3} c^2 a(t)^2 = 0$$

with a positive Λ , hence

$$\frac{da}{a} = \pm c \sqrt{\frac{\Lambda}{3}} dt, \quad \ln a - \ln a_0 = \pm c \sqrt{\frac{\Lambda}{3}} t$$

with a positive integration constant a_0 which is usually chosen as $a_0 = \sqrt{3/\Lambda}$. We only consider the solution with the plus sign. The solution with the minus sign gives the same universe with the time direction reversed. The scale factor

$$a(t) = \sqrt{\frac{3}{\Lambda}} \exp\left(c \sqrt{\frac{\Lambda}{3}} t\right) \quad (138)$$

expands monotonically from $t = -\infty$ to $t = +\infty$. The metric reads

$$g = -c^2 dt^2 + \frac{3}{\Lambda} \exp\left(\sqrt{\frac{\Lambda}{3}} 2ct\right) \left(d\chi^2 + \chi^2 (d\vartheta^2 + \sin^2\vartheta d\varphi^2)\right). \quad (139)$$

This is again the deSitter universe, but this time only half of it and with a slicing into *flat* 3-spaces. We see that a different choice of the integration constant a_0 could be compensated for by a rescaling of χ .

Note that this particular scale factor gives a “Hubble constant” that is really a constant, i.e., independent of time,

$$\frac{a'(t)}{a(t)} = c \sqrt{\frac{\Lambda}{3}}, \quad (140)$$

so the universe expands at a constant rate. The embedding into the full deSitter universe is shown in Figure 16.

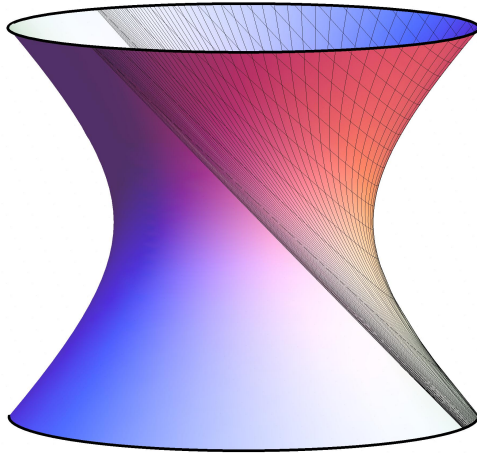


Figure 16: The deSitter universe with the slicing into flat hypersurfaces that covers half of the hyperboloid

The 3-dimensional flat slices $t = \text{constant}$ are represented as lines that come about as the sections of the hyperboloid with planes under 45 degrees. The boundary of the region covered corresponds to $t = -\infty$. The embedding is given by the equations

$$X = \sqrt{\frac{3}{\Lambda}} \exp\left(\sqrt{\frac{\Lambda}{3}} ct\right) \chi \cos \varphi \sin \vartheta, \quad (141)$$

$$Y = \sqrt{\frac{3}{\Lambda}} \exp\left(\sqrt{\frac{\Lambda}{3}} ct\right) \chi \sin \varphi \sin \vartheta, \quad (142)$$

$$Z = \sqrt{\frac{3}{\Lambda}} \exp\left(\sqrt{\frac{\Lambda}{3}} ct\right) \chi \cos \vartheta, \quad (143)$$

$$W = \sqrt{\frac{3}{\Lambda}} \cosh\left(\sqrt{\frac{\Lambda}{3}} ct\right) - \sqrt{\frac{3}{\Lambda}} \frac{\chi^2}{2} \exp\left(\sqrt{\frac{\Lambda}{3}} ct\right), \quad (144)$$

$$V = \sqrt{\frac{3}{\Lambda}} \sinh\left(\sqrt{\frac{\Lambda}{3}} ct\right) + \sqrt{\frac{3}{\Lambda}} \frac{\chi^2}{2} \exp\left(\sqrt{\frac{\Lambda}{3}} ct\right). \quad (145)$$

This spacetime model (i.e., half of the deSitter universe with a flat slicing) was suggested by Bondi, Gold and Hoyle as an alternative to the big-bang model. Because it exists forever, with a constant expansion rate, they called it the *steady-state universe*. In the steady-state theory there is no cosmological constant but rather a hypothetical ‘‘C-field’’ which is associated with a creation of matter out of nothing. The steady-state theory was a strong competitor of the big-bang theory until the discovery of the cosmic background radiation gave convincing evidence for the latter.

- $k = -1$

In this case the Friedmann vacuum equation (121) reads

$$-c^2 + a'(t)^2 - \frac{\Lambda}{3} c^2 a(t)^2 = 0,$$

$$\frac{da}{dt} = \pm c \sqrt{\frac{\Lambda}{3} a^2 + 1}.$$

As Λ is positive, we may substitute

$$\sqrt{\frac{\Lambda}{3}} a = \sinh u, \quad \sqrt{\frac{\Lambda}{3}} da = \cosh u du,$$

hence

$$\sqrt{\frac{3}{\Lambda}} \frac{\cosh u du}{\sqrt{\sinh^2 u + 1}} = \pm c dt.$$

This results in

$$u = \pm \sqrt{\frac{\Lambda}{3}} c(t - t_0),$$

$$a(t) = \pm \sqrt{\frac{3}{\Lambda}} \sinh\left(\sqrt{\frac{\Lambda}{3}} c(t - t_0)\right). \quad (146)$$

We choose $t_0 = 0$ and the plus sign. (Again, the minus sign gives a time-reversed version of the same spacetime.) The scale factor is defined for $0 < t < \infty$ and increases monotonically. The metric reads

$$g = -c^2 dt^2 + \frac{3}{\Lambda} \sinh^2 \left(\sqrt{\frac{\Lambda}{3}} ct \right) \left(d\chi^2 + \sinh^2 \chi (d\vartheta^2 + \sin^2 \vartheta d\varphi^2) \right). \quad (147)$$

This is again part of the deSitter universe, this time with a slicing into hyperbolic spaces. The part covered by the slicing sits in the entire deSitter universe in a similar fashion as the Milne universe sits in the entire Minkowski spacetime, see Figure 17.

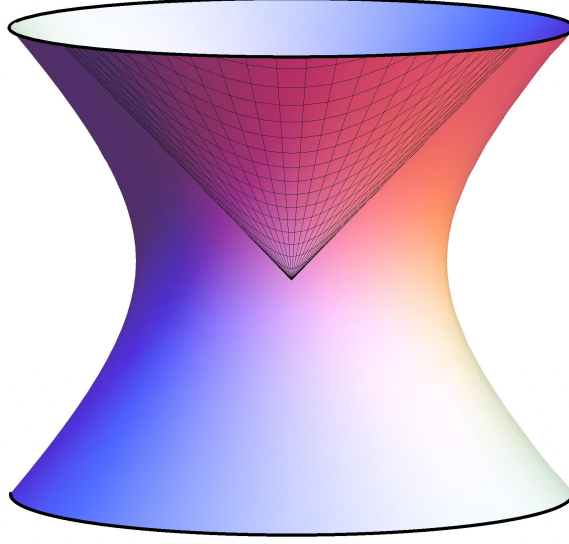


Figure 17: The deSitter universe with the slicing into hypersurfaces of constant negative curvature that covers the interior of the future null-cone of one event

The embedding into the hyperboloid is given by the equations

$$X = \sqrt{\frac{3}{\Lambda}} \sinh \left(\sqrt{\frac{\Lambda}{3}} ct \right) \sinh \chi \cos \varphi \sin \vartheta, \quad (148)$$

$$Y = \sqrt{\frac{3}{\Lambda}} \sinh \left(\sqrt{\frac{\Lambda}{3}} ct \right) \sinh \chi \sin \varphi \sin \vartheta, \quad (149)$$

$$Z = \sqrt{\frac{3}{\Lambda}} \sinh \left(\sqrt{\frac{\Lambda}{3}} ct \right) \sinh \chi \cos \vartheta, \quad (150)$$

$$W = \sqrt{\frac{3}{\Lambda}} \cosh \left(\sqrt{\frac{\Lambda}{3}} ct \right), \quad (151)$$

$$V = \sqrt{\frac{3}{\Lambda}} \sinh \left(\sqrt{\frac{\Lambda}{3}} ct \right) \cosh \chi. \quad (152)$$

(iii) $\Lambda < 0$

- $k = 1$

In this case there is no solution because the differential equation

$$c^2 + a'(t)^2 - \frac{\Lambda}{3} c^2 a(t)^2 = 0,$$

with a negative Λ cannot be satisfied by a real function a .

- $k = 0$

Again, there is no solution because

$$a'(t)^2 - \frac{\Lambda}{3} c^2 a(t)^2 = 0,$$

cannot hold for a real a if Λ is negative.

- $k = -1$

In this case the Friedmann vacuum equation (121) reads

$$-c^2 + a'(t)^2 - \frac{\Lambda}{3} c^2 a(t)^2 = 0$$

which requires

$$\frac{da}{dt} = \pm c \sqrt{1 + \frac{\Lambda}{3} a^2}.$$

As Λ is negative, we may substitute

$$\sqrt{-\frac{\Lambda}{3}} a = \sin u,$$

$$\sqrt{\frac{-\Lambda}{3}} da = \cos u du$$

hence

$$\sqrt{-\frac{3}{\Lambda}} \frac{\cos u du}{\sqrt{1 - \sin^2 u}} = \pm c dt.$$

This results in

$$u - u_0 = \pm \sqrt{\frac{\Lambda}{3}} c t$$

with an integration constant u_0 . If we choose $u_0 = 0$ and the plus sign, the scale factor

$$a(t) = \sqrt{-\frac{3}{\Lambda}} \sin\left(\sqrt{\frac{-\Lambda}{3}} c t\right) \quad (153)$$

is defined on the time interval $0 < t < (-3/\Lambda)^{-1/2} \pi/c$. Here the other choice of the sign (and another choice of the integration constant) gives the same behaviour because the situation is time symmetric. The universe starts with a “big bang” and ends in a “big crunch”. The metric reads

$$g = -c^2 dt^2 - \frac{3}{\Lambda} \sin^2\left(\sqrt{-\frac{\Lambda}{3}} c t\right) \left(d\chi^2 + \sinh^2 \chi (d\vartheta^2 + \sin^2 \vartheta d\varphi^2)\right). \quad (154)$$

This spacetime is part of the so-called *anti-deSitter universe*. The full anti-deSitter universe is the isometrically embedded hyperboloid

$$X^2 + Y^2 + Z^2 - W^2 - V^2 = -\frac{3}{\Lambda} \quad (155)$$

in the 5-dimensional pseudo-Euclidean space with metric

$$g^{(5)} = dX^2 + dY^2 + dZ^2 - dW^2 - dV^2. \quad (156)$$

The full anti-deSitter universe has the topology of $\mathbb{R}^3 \times S^1$ where S^1 is a 1-dimensional sphere, i.e., a circle. The cyclic dimension is timelike, i.e., in the anti-deSitter universe there are closed timelike curves through each point. One can remove this unwanted feature by considering the universal covering space.

The slicing into hyperbolic spaces covers only part of the anti-deSitter universe, see Figure 18. The embedding into the hyperboloid is given by the equations

$$X = \sqrt{\frac{3}{\Lambda}} \sin\left(\sqrt{\frac{\Lambda}{3}} ct\right) \sinh \chi \cos \varphi \sin \vartheta, \quad (157)$$

$$Y = \sqrt{\frac{3}{\Lambda}} \sin\left(\sqrt{\frac{\Lambda}{3}} ct\right) \sinh \chi \sin \varphi \sin \vartheta, \quad (158)$$

$$Z = \sqrt{\frac{3}{\Lambda}} \sin\left(\sqrt{\frac{\Lambda}{3}} ct\right) \sinh \chi \cos \vartheta, \quad (159)$$

$$W = \sqrt{\frac{3}{\Lambda}} \cos\left(\sqrt{\frac{\Lambda}{3}} ct\right), \quad (160)$$

$$V = \sqrt{\frac{3}{\Lambda}} \sin\left(\sqrt{\frac{\Lambda}{3}} ct\right) \cosh \chi. \quad (161)$$

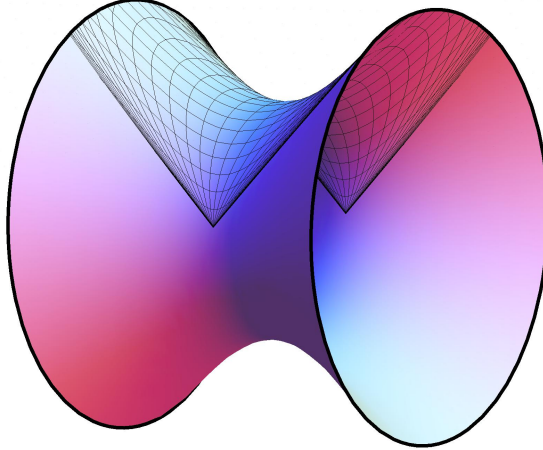


Figure 18: The anti-deSitter universe with the slicing into hypersurfaces of constant negative curvature that covers the interior of the future null-cone of one event and, at the same time, the interior of the past null-cone of another event

The anti-deSitter spacetime is not considered as a realistic model of our universe, not even in the sense of a limit. However, anti-deSitter (AdS) spaces of various dimension play an important role in string theory, in particular as mathematical tools for calculations in conformal field theories (CFT). This approach is known as the AdS-CFT correspondence.

We summarise our results on vacuum solutions to the Friedmann equations. We have seen that the only solutions are Minkowski spacetime ($\Lambda = 0$), deSitter spacetime ($\Lambda > 0$) and anti-deSitter spacetime ($\Lambda < 0$). Minkowski spacetime can be viewed as a Robertson-Walker spacetime in two different ways, with slices of curvature $k = 0$ or $k = -1$. For the deSitter spacetime all three kinds of slicings, $k = 1$, $k = 0$ and $k = -1$, are possible, whereas for the anti-deSitter spacetime it only works with $k = -1$. The slicings with $k = -1$ have an initial singularity in the sense that the standard observers are compressed into one point, but this is of course not a curvature singularity. To put this another way, it is a singularity of the slicing and not of the spacetime.

Minkowski, deSitter and anti-deSitter spacetime have constant curvature $\Lambda/3$, i.e., the curvature tensor satisfies

$$R_{\nu\sigma\tau\mu} = \frac{\Lambda}{3} (g_{\nu\tau}g_{\sigma\mu} - g_{\nu\mu}g_{\sigma\tau})$$

as can be verified in any of the given coordinate representations. Also, they are the only spacetimes with ten linearly independent Killing vector fields which is the maximal number in a pseudo-Riemannian manifold of dimension 4. (Here “linearly independent” refers to linear combinations with *constant* coefficients; of course, if we allow for coefficients that depend on the foot-point there cannot be more than four linearly independent vector fields.) In the Minkowski case, $\Lambda = 0$, these ten Killing vector fields generate the Poincaré group, i.e., the 4 translations, the 3 spatial rotations and the 3 Lorentz boosts. The corresponding symmetry groups for $\Lambda > 0$ and $\Lambda < 0$ are known as the *deSitter group* and the *anti-deSitter group*, respectively.

We have seen that Minkowski spacetime with the flat slicing is the only vacuum solution to the Friedmann equations with a constant scale factor. This does not mean that the deSitter spacetime and the anti-deSitter spacetime are not static: Actually, they do admit a timelike Killing vector field that is perpendicular to spacelike slices, but these slices are not spaces of constant curvature, so in this static representation the metric does not have the Robertson-Walker form. We will discuss these static representations in Worksheet 4.

(b) Dust solutions

We now consider dust solutions to the Friedmann equations (119) and (120), i.e., solutions with $p = 0$. Then there is no internal energy, i.e., $\varepsilon = \mu c^2$ where μ is the (rest-)mass density. A dust is a good mathematical model for the ordinary (baryonic) matter in galaxies and also for “cold” dark matter. Quite generally, the terms “cold fluid” and “dust” are synonymous, meaning a perfect fluid with vanishing pressure. However, this terminology can be misleading: E.g., the Solar Corona has a temperature of almost 10^7 Kelvins, so what would not call it “cold” by ordinary standards. Nonetheless the pressure in the Solar Corona is negligibly small because the density is rather low.

With (119) and (120) specified to the dust case we have to solve the equations

$$\frac{3c^2k}{c^2a(t)^2} + \frac{3a'(t)^2}{c^2a(t)^2} - \Lambda = \kappa c^2 \mu(t), \quad (162)$$

$$-k - \frac{1}{c^2} a'(t)^2 - 2 \frac{a(t)}{c^2} a''(t) + \Lambda a(t)^2 = 0. \quad (163)$$

We require $\mu(t) > 0$ throughout.

We first look for static solutions, $a(t) = a_0 = \text{constant}$. Then (162) and (163) require

$$\begin{aligned} \frac{3c^2k}{a_0^2} - \Lambda c^2 &= \kappa c^4 \mu(t), \\ -k + \Lambda a_0^2 &= 0. \end{aligned}$$

As expected, this can be true only if the density is constant, $\mu(t) = \mu_0$. Solving for Λ and μ_0 yields

$$\Lambda = \frac{k}{a_0^2}, \quad \mu_0 = \frac{2k}{\kappa c^2 a_0^2}. \quad (164)$$

As we assume that the density is positive, the second equation implies $k = 1$, so the first one requires $\Lambda > 0$. We summarise this important result in the following way:

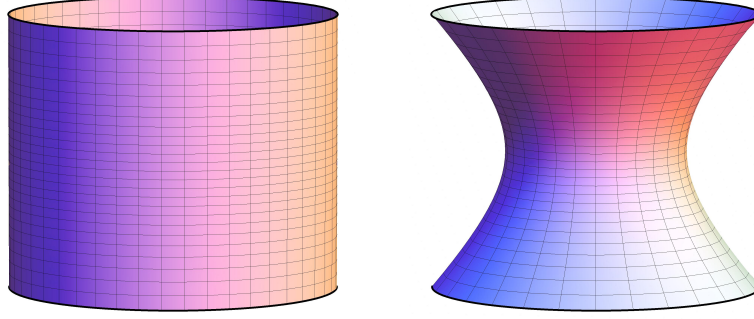
If we consider Einstein's field equation without a cosmological constant, there is no static solution to the Friedmann equations for a dust of positive mass density.

It was this observation that led Einstein to introducing the cosmological constant in 1917. With a positive cosmological constant, a static dust solution does exist. It has $k = 1$, so the natural spatial topology is that of a 3-sphere. The metric reads

$$g = -c^2 dt^2 + a_0^2 \left(d\chi^2 + \sin^2 \chi (d\vartheta^2 + \sin^2 \vartheta d\varphi^2) \right) \quad (165)$$

where a_0 is a positive constant. By (164), the corresponding cosmological constant is $\Lambda = 1/a_0^2$ and the mass density is $\mu_0 = 2/(\kappa c^2 a_0^2)$.

This is *Einstein's static universe*, also known as *Einstein's cylinder spacetime*, referring to the fact that the natural topology of this spacetime is $S^3 \times \mathbb{R}$ with a 3-sphere S^3 of constant radius a_0 . Einstein advertised this spacetime as a viable mathematical model of our universe in 1917. In the same year, deSitter introduced the universe named after him, recall eq. (127). Until Friedmann's work in 1922/23, these two spacetimes were the only cosmological models that were discussed on the basis of general relativity. We compare them in the following table, where we consider the “natural” (global) slicing of the deSitter universe as it is given in eq. (127).



	Einstein's static universe	deSitter universe
topology	$S^3 \times \mathbb{R}$	$S^3 \times \mathbb{R}$
time dependence	static	contracting, then expanding
Lambda term	$\Lambda > 0$	$\Lambda > 0$
matter	$\varepsilon = c^2 \mu > 0, p = 0$	$\varepsilon = 0, p = 0$

We see that a positive cosmological constant has a repellent effect. In Einstein's static universe this repellent effect is balanced by the gravitational attraction of the dust. In the deSitter universe the cosmological constant decelerates the initial contraction and then causes a re-expansion. Both spacetime models are eternal, without a big bang and without a big crunch: The scale factor $a(t)$ is defined and positive for all $t \in \mathbb{R}$.

Having found all static solutions to the equations (162) and (163), we assume $a'(t) \neq 0$ from now on. We first demonstrate in the dust case the two Friedmann equations imply a conservation law.

Claim: (162) and (163) imply the conservation law

$$\mu(t)a(t)^3 = \text{constant}. \quad (166)$$

Proof: We write (162) in the form

$$\frac{\kappa}{3} c^2 \mu(t) a(t)^3 = k a(t) + \frac{a(t)}{c^2} a'(t)^2 - \frac{\Lambda}{3} a(t)^3$$

and differentiate with respect to t :

$$\begin{aligned} \frac{\kappa}{3} c^2 \frac{d(\mu(t) a(t)^3)}{dt} &= k a'(t) + \frac{a'(t)^3}{c^2} + \frac{2}{c^2} a(t) a'(t) a''(t) - \Lambda a(t)^2 a'(t) \\ &= a'(t) \left(k + \frac{a'(t)^2}{c^2} + \frac{2}{c^2} a(t) a''(t) - \Lambda a(t)^2 \right). \end{aligned}$$

By (163), this is indeed zero.

Of course, this result just establishes the conservation of mass. (For a dust, the only form of rest energy is the rest mass of the dust particles.) As the equation $\nabla^\mu T_{\mu\nu} = 0$ is a consequence of Einstein's field equation, it should not come as a surprise that this conservation law is implied by the Friedmann equations.

Hence, the most convenient way of solving the Friedmann equations for dust is by determining $a(t)$ from the differential equation

$$k a(t) + \frac{a(t)}{c^2} a'(t)^2 - \frac{\Lambda}{3} a(t)^3 = a_0, \quad (167)$$

with a positive constant a_0 which must have the dimension of a length. Once a solution $a(t)$ has been found, the corresponding density $\mu(t)$ is determined by the equation

$$\frac{\kappa}{3} c^2 \mu(t) a(t)^3 = a_0. \quad (168)$$

We will now solve (167) for all values of k with $\Lambda = 0$. Thereafter, we will discuss how a non-vanishing Λ influences these solutions.

(i) $\Lambda = 0$

Then (167) reduces to

$$k a + \frac{a}{c^2} \left(\frac{da}{dt} \right)^2 = a_0. \quad (169)$$

For solving this differential equation it is convenient to use the conformal time η . Recall that, by (77), the latter satisfies

$$d\eta = \frac{c dt}{a}, \quad (170)$$

hence

$$\frac{da}{dt} = \frac{da}{d\eta} \frac{d\eta}{dt} = \frac{da}{d\eta} \frac{c}{a}.$$

Then the differential equation takes the form

$$k a + \frac{a}{c^2} \frac{c^2}{a^2} \left(\frac{da}{d\eta} \right)^2 = a_0,$$

$$\left(\frac{da}{d\eta} \right)^2 = a_0 a - k a^2.$$

As we know that $a(t) = \text{constant}$ cannot be a solution with $\Lambda = 0$, we may assume that the left-hand side and, thus, the right-hand side is different from zero, hence

$$\frac{da}{\sqrt{a_0 a - k a^2}} = \pm d\eta. \quad (171)$$

• $k = 1$

Then we have to integrate

$$\frac{2 da}{a_0 \sqrt{1 - \left(1 - \frac{2a}{a_0}\right)^2}} = \pm d\eta.$$

The integration can be done with a substitution,

$$1 - \frac{2a}{a_0} = \cos u, \quad \frac{2 da}{a_0} = \sin u du.$$

The resulting integral reads

$$\int \frac{\sin u \, du}{\sqrt{1 - \sin^2 u}} = \pm \int d\eta,$$

$$u - u_0 = \pm \eta$$

with an integration constant u_0 . The choice of the integration constant is irrelevant because it has the only effect of shifting the zero on the time axis. We choose $u_0 = 0$. This yields

$$1 - \frac{2a}{a_0} = \cos \eta,$$

$$a = \frac{a_0}{2} (1 - \cos \eta) = \hat{a}(\eta). \quad (172)$$

We have thus determined the scale factor as a function of the conformal time η . With this result, the relation (170) between t and η reads

$$c \, dt = a \, d\eta = \frac{a_0}{2} (1 - \cos \eta) d\eta,$$

and, upon choosing the integration constant appropriately,

$$t = \frac{a_0}{2c} (\eta - \sin \eta) = \hat{t}(\eta). \quad (173)$$

(172) and (173) give us the graph of the function $t \mapsto a$ in parametric form, with η as the curve parameter. The resulting curve is a cycloid, see Figure 19. (A cycloid is the curve traced by a point on the rim of a wheel that is rolling on a horizontal surface along a straight line.) From this figure we read that the universe begins with a big bang at $t = 0$, reaches a maximal extension at $t = a_0\pi/(2c)$ and ends in a big crunch at $t = a_0\pi/c$. By (168), the mass density is infinite at the big bang, decreases to a minimum that is reached at $t = a_0\pi/(2c)$ and then increases to infinity at the big crunch.

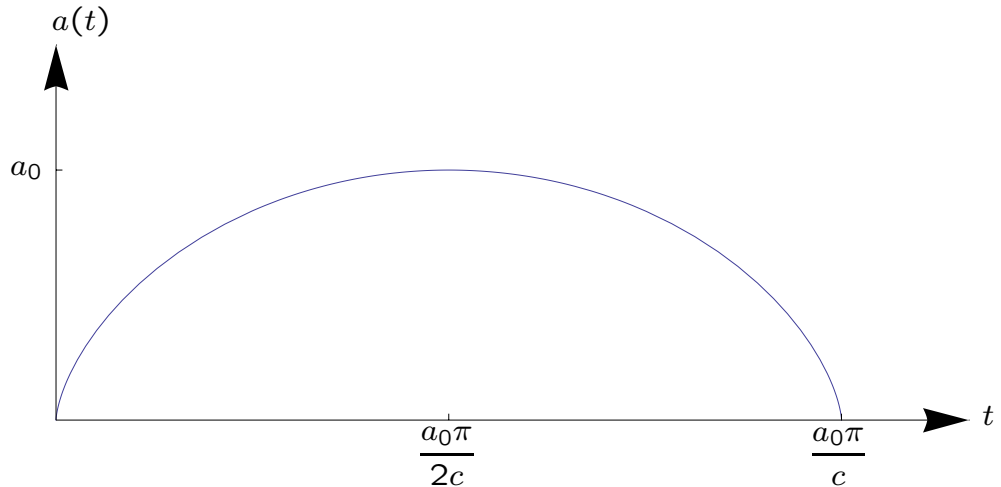


Figure 19: Scale factor in a dust universe with $k = +1$

- $k = 0$

In this case it is not actually necessary to consider the conformal time, but in view of consistency with the other cases we proceed analogously. Integrating (171) for $k = 0$,

$$\frac{1}{\sqrt{a_0}} \int \frac{da}{\sqrt{a}} = \pm \int d\eta$$

yields

$$\frac{2}{\sqrt{a_0}} \sqrt{a} - C_0 = \pm \eta$$

with an integration constant C_0 . Again, the choice of the integration constant produces just a shift of the zero on the time axis. We choose $C_0 = 0$. Then

$$\frac{4a}{a_0} = \eta^2.$$

The relation (170) between t and η reads

$$c dt = a d\eta = \frac{a_0}{4} \eta^2 d\eta,$$

and upon integration

$$c t = \frac{a_0}{4} \frac{\eta^3}{3};$$

here we have chosen the initial condition such that $t = 0$ corresponds to $\eta = 0$. Hence

$$a(t) = \frac{a_0}{4} \eta^2 = \frac{a_0}{4} \left(\frac{12 c t}{a_0} \right)^{2/3} = \left(\frac{9}{4} a_0 c^2 \right)^{1/3} t^{2/3}. \quad (174)$$

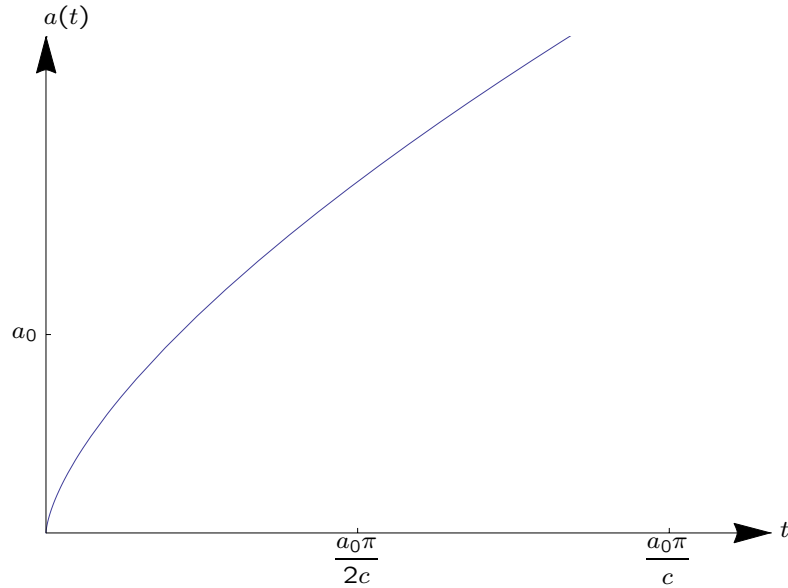


Figure 20: Scale factor in a dust universe with $k = 0$ (Einstein-deSitter universe)

The metric reads

$$g = -c^2 dt^2 + \left(\frac{9}{4} a_0 c^2\right)^{2/3} t^{4/3} \left(d\chi^2 + \chi^2 (d\vartheta^2 + \sin^2 \vartheta d\varphi^2)\right). \quad (175)$$

This is the *Einstein-deSitter universe*. It is a dust-filled, forever expanding spacetime model with flat spatial sections, so the natural topology of this spacetime is \mathbb{R}^4 . The expansion is decelerated, as can be seen from (174) and from Figure 20. In Worksheet 3 we will demonstrate that the deceleration parameter is indeed positive and that it is actually a constant, $q(t_o) = 1/2$ for all $0 < t_o < \infty$.

After the idea of an expanding universe had been widely accepted around 1930, Einstein and deSitter advertised this special solution to the Friedmann equations as the most promising cosmological world model in a joint paper. It was considered as the best model for the present state of our universe until in the late 1990s the supernova Ia observations gave strong evidence for an *accelerated* expansion, i.e., for a negative value of $q(t_o)$.

- $k = -1$

The mathematics is quite similar to the case $k = 1$. We have to integrate

$$\frac{2 da}{a_0 \sqrt{\left(\frac{2a}{a_0} + 1\right)^2 - 1}} = \pm d\eta$$

which can be done with the substitution

$$\frac{2a}{a_0} + 1 = \cosh u, \quad \frac{2 da}{a_0} = -\sinh u du.$$

Then the integral reads

$$\int \frac{\sinh u du}{\sqrt{\cosh^2 u - 1}} = \pm \int d\eta,$$

$$u - u_0 = \pm \eta$$

where we choose the integration constant u_0 again equal to zero. This yields

$$\operatorname{arcosh}\left(\frac{2a}{a_0} + 1\right) = \pm \eta,$$

$$\frac{2a}{a_0} + 1 = \cosh(\eta),$$

$$a = \frac{a_0}{2} (\cosh(\eta) - 1) = \hat{a}(\eta).$$

With this result, the relation (170) between t and η reads

$$c dt = a d\eta = \frac{a_0}{2} (\cosh \eta - 1) d\eta,$$

and, upon choosing the integration constant appropriately,

$$t = \frac{a_0}{2c} (\sinh \eta - \eta) = \hat{t}(\eta).$$

Again, we have found the graph of the function $t \mapsto a$ in parametric form, with η as the curve parameter. The resulting curve is the hyperbolic analogue of a cycloid, sometimes called a “hyperbolic cycloid”. The universe begins with a big bang at $t = 0$ and expands forever, see Figure 21.

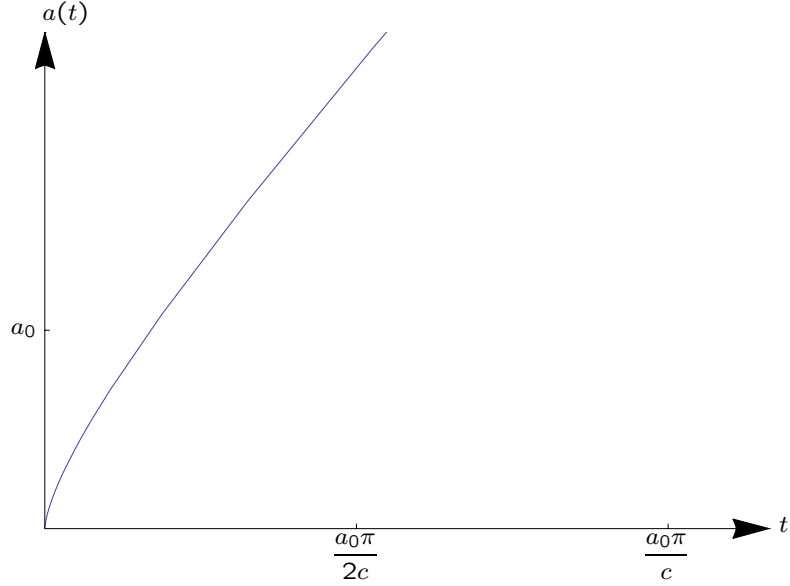


Figure 21: Scale factor in a dust universe with $k = -1$

We see that, for $k = 1$, $k = 0$ and $k = -1$, the dust universe without a cosmological constant is always *decelerating*, i.e., $q(t_o) > 0$ for all t_o . For $k = -1$, the initial “explosion” is strong enough to make the universe expand forever. For $k = 1$, however, the self-gravitating dust is dense enough to make the universe re-collapse into a big crunch. The case $k = 0$ is the critical case where “the turning point is at infinity”, i.e., the universe just makes it to avoid re-collapse. This borderline case can be characterised by a critical density in the following way. At any chosen time t_o , the first Friedmann equation without a cosmological constant for a dust, i.e., eq. (162), can be written as an equality between densities,

$$\frac{3c^2k}{\kappa c^4 a(t_o)^2} = \mu(t_o) - \frac{3a'(t_o)^2}{\kappa c^4 a(t_o)^2}. \quad (176)$$

The sign of the left-hand side is determined by k . If we define a critical density (which depends on t_o) by

$$\mu_c(t_o) := \frac{3a'(t_o)^2}{\kappa c^4 a(t_o)^2} = \frac{3H(t_o)^2}{\kappa c^4}, \quad (177)$$

we see that

$$\begin{aligned}\mu(t_o) &> \mu_c(t_o) & \text{if } k = 1, \\ \mu(t_o) &= \mu_c(t_o) & \text{if } k = 0, \\ \mu(t_o) &< \mu_c(t_o) & \text{if } k = -1.\end{aligned}$$

One often uses the *density parameter*

$$\Omega_m(t_o) = \frac{\mu(t_o)}{\mu_c(t_o)} \quad (178)$$

where the index m (for “matter”) is meant as a reminder that we are considering a dust universe here. Figure 22 shows the behaviour of the scale factor for the three cases in one diagram: Solid for overcritical density, dashed for critical density and dotted for undercritical density.

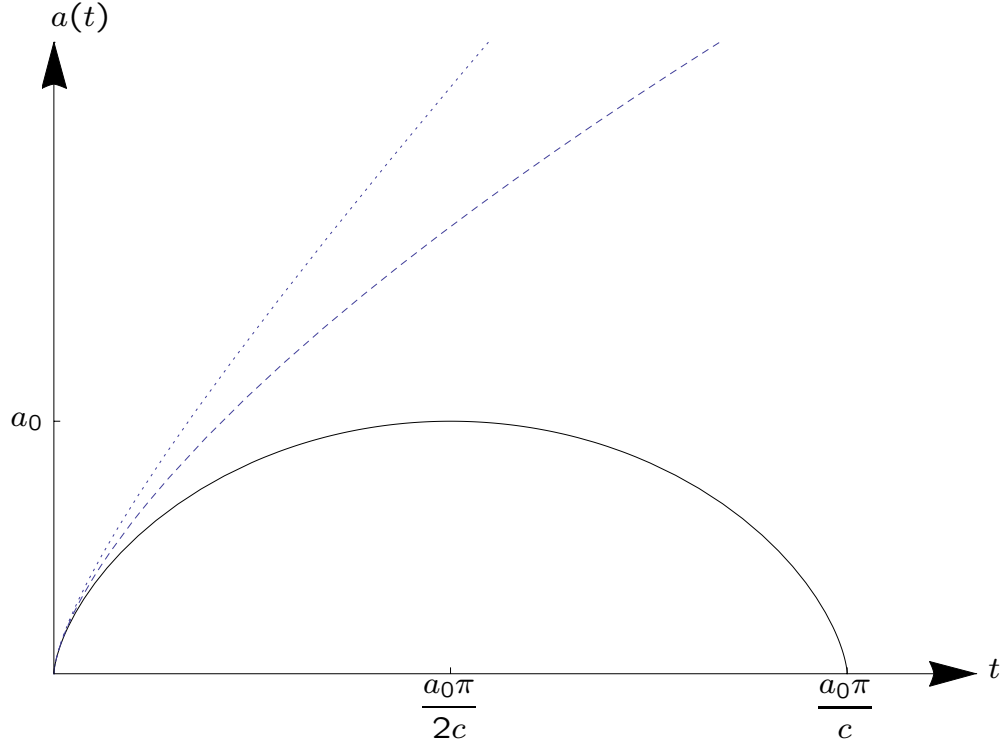


Figure 22: Dust universes without a cosmological constant

For the dust universes without a cosmological constant W. Mattig found in 1957 an *exact* distance-redshift relation which involved only the parameters $H(t_o)$ and $q(t_o)$. This is quite remarkable because in general $H(t_o)$ and $q(t_o)$ determine the distance-redshift relations only up to second order with the higher-order terms completely independent. The Mattig formula reads

$$D_L = \frac{c}{H(t_o) q(t_o)^2} \left\{ q(t_o) z + \left(q(t_o) - 1 \right) \left(\sqrt{2 q(t_o) z + 1} - 1 \right) \right\}. \quad (179)$$

Here D_L is the luminosity distance. This formula is true in all three types of dust universes without a cosmological constant, i.e., for $k = 1$, $k = 0$ and $k = -1$. However, in the case $k = 1$ one has to restrict to observation times $0 < t_o < a_0\pi/(2c)$ where $a'(t_o) > 0$, see the proof of the Mattig formula below. For later times, where $a'(t_o) < 0$, the relation between z and D_L is no longer one-to-one: It can be shown that then there are two values of D_L corresponding to the same value of z and only one of them is given by the Mattig formula. We will now prove the Mattig formula which is elementary but a bit cumbersome.

Proof of the Mattig formula: In the following we have to assume that $a'(t) > 0$. If $k = -1$ or $k = 0$, this is true for all $t > 0$; if $k = 1$, however, this assumption restricts the validity to times $0 < t < a_0\pi/(2c)$. As a preparation, we first demonstrate that the following three auxiliary equations hold.

$$a'(t) = c \sqrt{\frac{a_0}{a(t)} - k}, \quad (180)$$

$$k \frac{a(t_o)}{a_0} = 1 - \frac{1}{2q(t_o)}. \quad (181)$$

$$a(t_o) = \frac{c\sqrt{k}}{H(t_o)\sqrt{2q(t_o) - 1}}. \quad (182)$$

(180) follows immediately by solving the Friedmann equation (169) for $a'(t)$, assuming that $a'(t) > 0$. To prove (181), we divide (169) by $a(t)$ and differentiate with respect to t . This results in

$$2a'(t)a''(t) = -\frac{c^2a_0}{a(t)^2}a'(t).$$

Assuming $a'(t) > 0$ we may multiply with $a(t)/a'(t)^3$,

$$\frac{2a(t)a''(t)}{a'(t)^2} = -\frac{c^2a_0}{a(t)a'(t)^2},$$

which, by (180), can be rewritten as

$$-\frac{2a(t)a''(t)}{a'(t)^2} = \frac{a_0}{a(t)\left(\frac{a_0}{a(t)} - k\right)}.$$

Evaluation at time t_o gives the deceleration parameter at this time,

$$2q(t_o) = \frac{a_0}{a_0 - k a(t_o)},$$

hence (181). The third auxiliary equation (182) follows from inserting (180) into the equation that defines the Hubble constant,

$$H(t_o) = \frac{a'(t_o)}{a(t_o)} = \frac{c}{a(t_o)} \sqrt{\frac{a_0}{a(t_o)} - k} = \frac{c\sqrt{k}}{a(t_o)} \sqrt{\frac{a_0}{k a(t_o)} - 1}.$$

For $k = -1$, both square roots are purely imaginary, so the right-hand side is real. For $k = 0$, the right-hand side is to be understood in the sense of a limiting procedure. With (181), we find from the last equation

$$\begin{aligned} H(t_o) &= \frac{c\sqrt{k}}{a(t_o)} \sqrt{\frac{1}{1 - \frac{1}{2q(t_o)}} - 1} = \frac{c\sqrt{k}}{a(t_o)} \sqrt{\frac{\mathcal{I} - \mathcal{I} + \frac{1}{2q(t_o)}}{1 - \frac{1}{2q(t_o)}}} \\ &= \frac{c\sqrt{k}}{a(t_o) \sqrt{2q(t_o) - 1}}, \end{aligned}$$

hence (182). – It is now our goal to determine the luminosity distance D_L as a function of the redshift z . Recall from (102) that, for light from an emission event at time t_e to an observation event at time t_o , the luminosity distance is given by the equation

$$D_L = a(t_o) (1 + z) \xi(\chi)$$

where

$$\xi(\chi) = \frac{\sin(\sqrt{k}\chi)}{\sqrt{k}} = \begin{cases} \sin \chi & \text{if } k = 1, \\ \chi & \text{if } k = 0, \\ \sinh \chi & \text{if } k = -1. \end{cases}$$

Keeping the observation event t_o fixed, we shall determine $\xi(\chi)$ as a function of z . From

$$\chi = \int_{t_e}^{t_o} \frac{c dt}{a} = \int_{a(t_e)}^{a(t_o)} \frac{c dt da}{a da}$$

we find, with (180),

$$\chi = \int_{a(t_e)}^{a(t_o)} \frac{c da}{a \not{c} \sqrt{\frac{a_0}{a} - k}} = \int_{a(t_e)}^{a(t_o)} \frac{da}{\sqrt{a_o a - k a^2}}.$$

This is an elementary integral,

$$\begin{aligned} \chi &= \frac{1}{\sqrt{k}} \arcsin\left(\frac{2ka}{a_o} - 1\right) \Big|_{a(t_e)}^{a(t_o)}, \\ \sqrt{k}\chi &= \arcsin\left(\frac{2ka(t_o)}{a_o} - 1\right) - \arcsin\left(\frac{2ka(t_e)}{a_o} - 1\right) \\ &= \underbrace{\arcsin\left(\frac{2ka(t_o)}{a_o} - 1\right)}_{=\alpha} - \underbrace{\arcsin\left(\frac{2ka(t_o)}{a_o(1+z)} - 1\right)}_{=\beta}. \end{aligned}$$

Hence

$$\begin{aligned}
D_L &= a(t_o)(1+z) \frac{\sin(\sqrt{k}\chi)}{\sqrt{k}} = a(t_o)(1+z) \frac{\sin(\alpha - \beta)}{\sqrt{k}} \\
&= \frac{a(t_o)}{\sqrt{k}} (1+z) \left(\sin \alpha \cos \beta - \sin \beta \cos \alpha \right) \\
&= \frac{a(t_o)}{\sqrt{k}} (1+z) \left(\sin \alpha \sqrt{1 - \sin^2 \beta} - \sin \beta \sqrt{1 - \sin^2 \alpha} \right).
\end{aligned}$$

Inserting the expressions for α and β yields

$$\begin{aligned}
D_L &= \frac{a(t_o)}{\sqrt{k}} (1+z) \left\{ \left(\frac{2ka(t_o)}{a_0} - 1 \right) \sqrt{1 - \left(\frac{2ka(t_o)}{a_0(1+z)} - 1 \right)^2} \right. \\
&\quad \left. - \left(\frac{2ka(t_o)}{a_0(1+z)} - 1 \right) \sqrt{1 - \left(\frac{2ka(t_o)}{a_0} - 1 \right)^2} \right\} \\
&= \frac{a(t_o)}{\sqrt{k}} \left\{ \left(\frac{2ka(t_o)}{a_0} - 1 \right) \sqrt{(1+z)^2 - \left(\frac{2ka(t_o)}{a_0} - 1 - z \right)^2} \right. \\
&\quad \left. - \left(\frac{2ka(t_o)}{a_0} - 1 - z \right) \sqrt{1 - \left(\frac{2ka(t_o)}{a_0} - 1 \right)^2} \right\}.
\end{aligned}$$

With (181) this can be rewritten as

$$\begin{aligned}
D_L &= \frac{a(t_o)}{\sqrt{k}} \left\{ \left(1 - \frac{1}{q(t_o)} \right) \sqrt{(1+z)^2 - \left(1 - z - \frac{1}{q(t_o)} \right)^2} \right. \\
&\quad \left. - \left(1 - z - \frac{1}{q(t_o)} \right) \sqrt{1 - \left(1 - \frac{1}{q(t_o)} \right)^2} \right\} \\
&= \frac{a(t_o)}{\sqrt{k} q(t_o)^2} \left\{ \left(q(t_o) - 1 \right) \sqrt{q(t_o)^2 (1+z)^2 - \left(q(t_o)(1-z) - 1 \right)^2} \right. \\
&\quad \left. - \left(q(t_o) - q(t_o)z - 1 \right) \sqrt{q(t_o)^2 - \left(q(t_o) - 1 \right)^2} \right\} \\
&= \frac{a(t_o)}{\sqrt{k} q(t_o)^2} \left\{ \left(q(t_o) - 1 \right) \sqrt{q(t_o)^2 4z + 2q(t_o)(1-z) - 1} \right. \\
&\quad \left. - \left(q(t_o) - q(t_o)z - 1 \right) \sqrt{2q(t_o) - 1} \right\} \\
&= \frac{a(t_o) \sqrt{2q(t_o) - 1}}{\sqrt{k} q(t_o)^2} \left\{ \left(q(t_o) - 1 \right) \sqrt{2q(t_o)z + 1} - q(t_o) + q(t_o)z + 1 \right\}.
\end{aligned}$$

Finally, with (182), we get the Mattig formula.

□

(ii) $\Lambda \neq 0$

The Friedmann equation (167) can be solved in the case $\Lambda \neq 0$ with the same method as in the case $\Lambda = 0$, by introducing the conformal time. In this case one finds the parametric relation between the scale factor and conformal time given by an elliptic integral. We will not work this out here but restrict to a qualitative analysis.

To that end we write (167) in the form

$$\left(\frac{da}{dt}\right)^2 - \frac{a_0 c^2}{a} - \frac{\Lambda}{3} c^2 a^2 = -k c^2. \quad (183)$$

This equation has the form of the energy-conservation law of classical mechanics for a particle moving in one spatial dimension,

$$\left(\frac{da}{dt}\right)^2 + V(a) = E. \quad (184)$$

In this analogy, we have to identify a with the position coordinate of the particle, usually denoted x , and

$$V(a) = -\frac{a_0 c^2}{a} - \frac{\Lambda}{3} c^2 a^2, \quad E = -k c^2. \quad (185)$$

As the “kinetic energy” $(da/dt)^2$ cannot be negative, we must have $V(a) \leq E$. Hence, for each value of E (i.e., for $k = 1$, $k = 0$ and $k = -1$) the accessible range of a is given by that part of the line $V(a) = E$ that lies above the graph of the potential. Points where $V(a) = E$ are turning points where da/dt is zero.

For $\Lambda = 0$ we can read from Figure 23 the following behaviour. If $k = -1$ (i.e., $E = c^2$), the universe starts with a big bang at $a = 0$ and expands forever up to infinity.

For $k = 1$ (i.e., $E = -c^2$) it starts with a big bang, reaches a maximum value where $da/dt = 0$ and then recollapses towards a big crunch. $k = 0$ is the borderline case where the turning point is at infinity, i.e., the universe just makes it to expand forever. These observations reproduce, in a qualitative fashion, what we have found with the exact analytical solutions for the case $\Lambda = 0$, recall Figure 22.

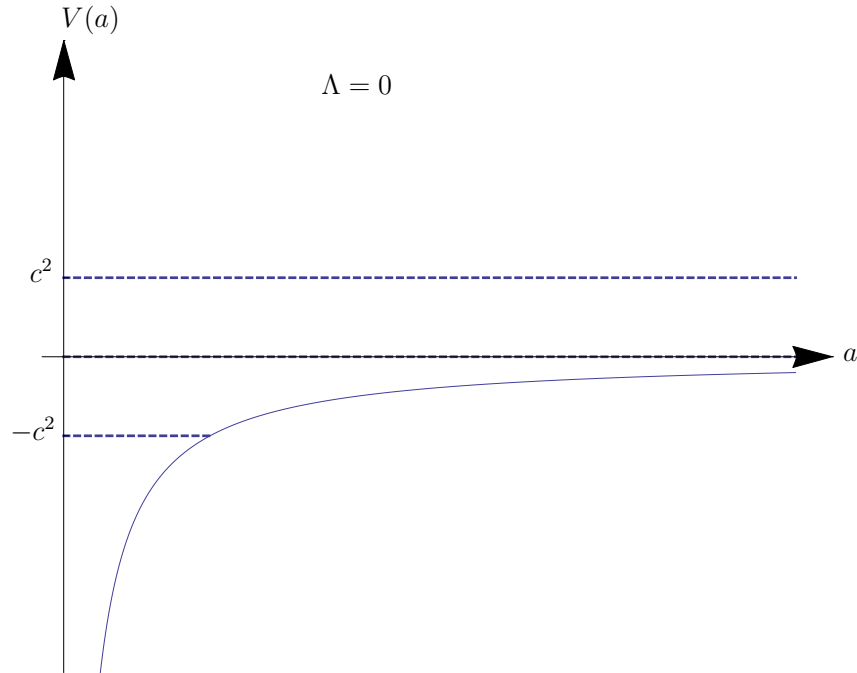


Figure 23: Potential $V(a)$ for $\Lambda = 0$

If $\Lambda < 0$, the potential $V(a)$ monotonically increases from $-\infty$ to ∞ . As a consequence, the universe is *always* recollapsing, for $k = -1$, $k = 0$ and $k = 1$. The maximal value a_{\max} of the scale factor is determined by the equation

$$V(a_{\max}) = -k c^2.$$

This is a cubic equation,

$$\frac{\Lambda}{3} c^2 a_{\max}^3 - k c^2 a_{\max} + a_0 c^2 = 0,$$

which has, indeed, precisely one real and positive solution a_{\max} if $\Lambda < 0$.

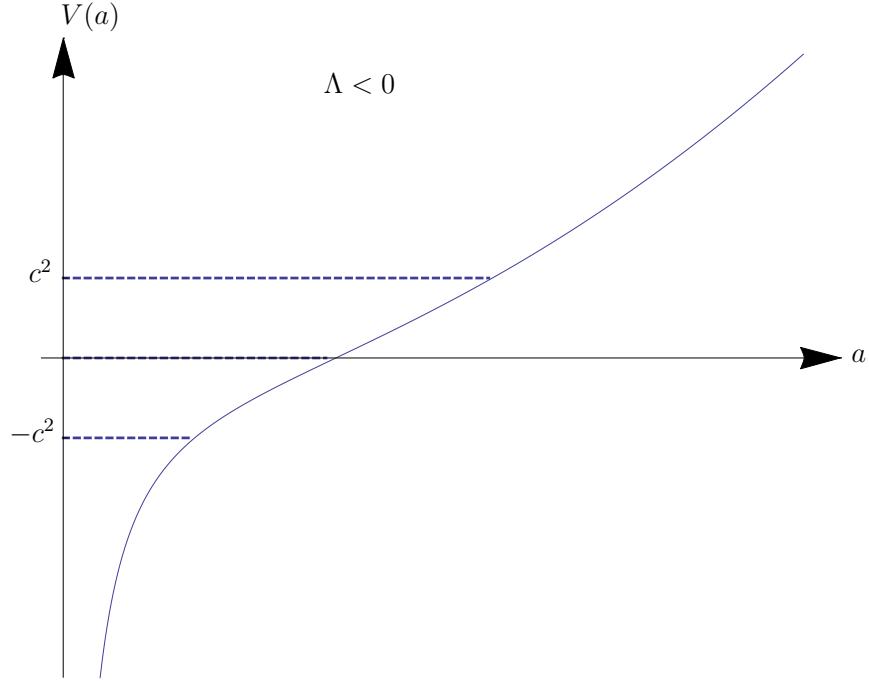


Figure 24: Potential $V(a)$ for $\Lambda < 0$

The case $\Lambda > 0$ is more subtle. Then the potential increases from $-\infty$ to a maximum at a certain value a_M and decreases again to $-\infty$. The behaviour of the universe with $k = 1$ depends on whether the maximum value $V(a_M)$ is smaller than, equal to or bigger than $-c^2$. The value a_M is determined by

$$V'(a_M) = \frac{a_0 c^2}{a_M^2} - \frac{2\Lambda}{3} c^2 a_M = 0, \quad a_M = \left(\frac{3a_0}{2\Lambda} \right)^{1/3}.$$

At this maximum, the potential takes the value

$$\begin{aligned} V(a_M) &= -a_0 c^2 \left(\frac{2\Lambda}{3a_0} \right)^{1/3} - \frac{\Lambda}{3} c^2 \left(\frac{3a_0}{2\Lambda} \right)^{2/3} \\ &= -a_0^{2/3} c^2 \left(\frac{2\Lambda}{3} \right)^{1/3} \left(1 + \frac{1}{2} \right) = -c^2 \left(\frac{9a_0^2}{4} \Lambda \right)^{1/3} \dots \end{aligned}$$

We see that

$$\begin{aligned} V(a_M) &> -c^2 & \text{if } \Lambda < \Lambda_{\text{crit}}, \\ V(a_M) &= -c^2 & \text{if } \Lambda = \Lambda_{\text{crit}}, \\ V(a_M) &< -c^2 & \text{if } \Lambda > \Lambda_{\text{crit}}, \end{aligned}$$

where $\Lambda_{\text{crit}} = \frac{4}{9a_0^2}$.

We consider first the case that $V(a_M) > -c^2$, i.e.

$$\Lambda < \frac{4}{9a_0^2} = \Lambda_{\text{crit}}.$$

If $k = -1$ or $k = 0$, the universe begins with a big bang and is expanding forever. In the case $k = 1$ there are two universes: One is starting with a big bang, reaches a maximum scale factor, and is then recollapsing. The other comes in from infinity, reaches a minimum scale factor and is then re-expanding to infinity.

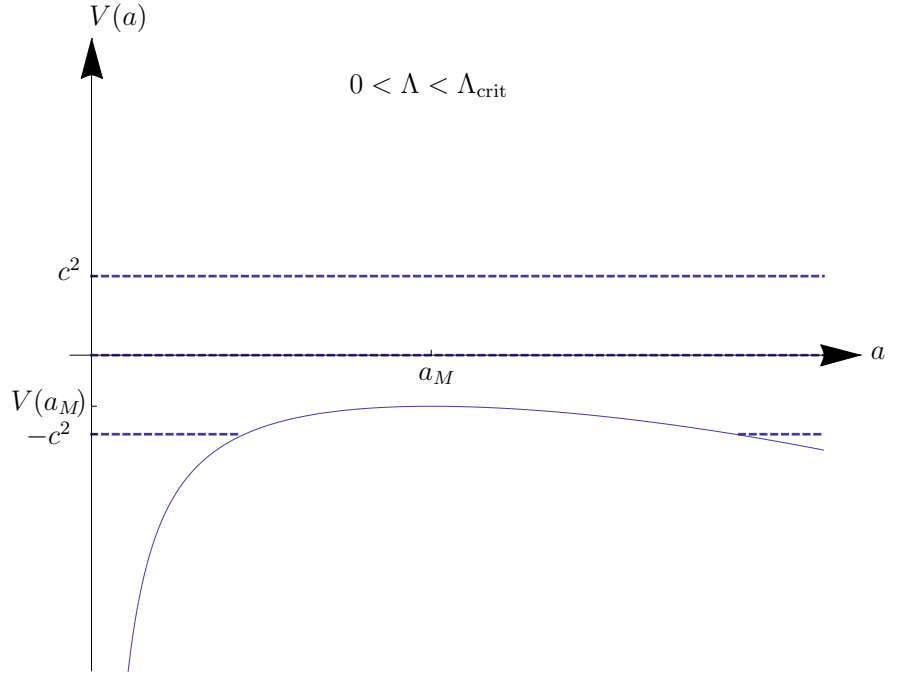


Figure 25: Potential $V(a)$ for $0 < \Lambda < \Lambda_{\text{crit}}$

If $V(a_M) = -c^2$, i.e.

$$\Lambda = \frac{4}{9a_0^2} = \Lambda_{\text{crit}},$$

for $k = -1$ or $k = 0$ the universe begins with a big bang and is expanding forever. In the case $k = 1$ there are two universes: One is starting with a big bang, then the scale factor is increasing and approaches the finite value a_M asymptotically from below. The other comes in from infinity, then the scale factor decreases monotonically and approaches the value a_M asymptotically from above.

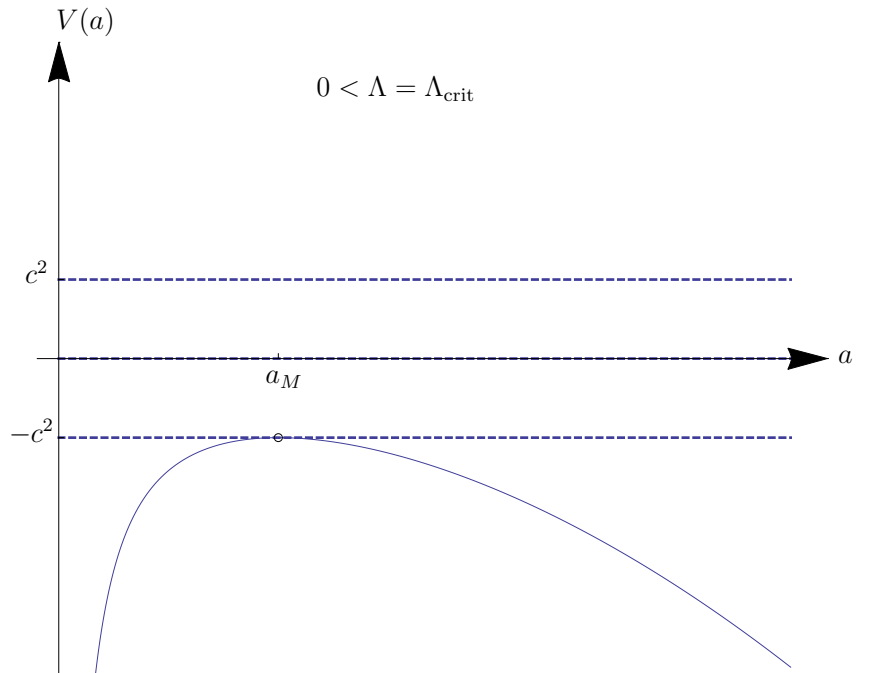


Figure 26: Potential $V(a)$ for $\Lambda = \Lambda_{\text{crit}}$

If $V(a_M) < -c^2$, i.e.

$$\Lambda > \frac{4}{9a_0^2} = \Lambda_{\text{crit}},$$

for all three cases $k = -1$, $k = 0$ and $k = 1$ the universe begins with a big bang and is expanding forever. From the differential equation (183) we read that a non-zero cosmological constant always dominates the dynamical behaviour for big a . This implies that, if $\Lambda > 0$, the spacetime is similar to the deSitter solution for big a .

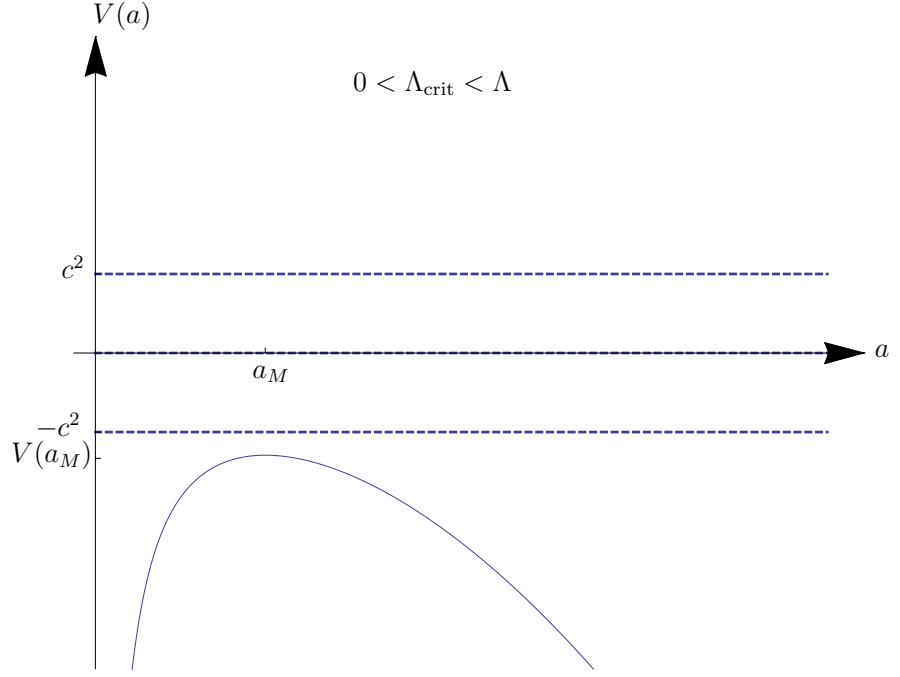


Figure 27: Potential $V(a)$ for $\Lambda_{\text{crit}} < \Lambda$

This is one of the reasons why the deSitter universe is relevant: Dust solutions with a positive cosmological constant asymptotically approach deSitter for $a \rightarrow \infty$.

(c) Perfect fluid solutions with pressure

We will now consider the Friedmann equations (119) and (120) in its generality, with a non-vanishing energy density ε and a non-vanishing pressure p ,

$$\frac{3c^2k}{a(t)^2} + \frac{3}{a(t)^2} a'(t)^2 - \Lambda c^2 = \kappa c^2 \varepsilon, \quad (186)$$

$$-\frac{k}{a(t)^2} - \frac{a'(t)^2}{c^2 a(t)^2} - \frac{2a''(t)}{c^2 a(t)} + \Lambda = \kappa p. \quad (187)$$

As long as we do not make any assumptions about ε or p , there is no equation to be solved: Given any k and any $a(t)$, one is even free to choose Λ at will and then (186) and (187) determine the energy density and the pressure. The equations are invariant under transformations

$$\Lambda \mapsto \Lambda - \Lambda_0, \quad \varepsilon(t) \mapsto \varepsilon(t) + \frac{\Lambda_0}{\kappa}, \quad p(t) \mapsto p(t) - \frac{\Lambda_0}{\kappa} \quad (188)$$

with an arbitrary real constant Λ_0 . We can utilise this fact for re-interpreting a solution with a cosmological constant as a solution without a cosmological constant, by choosing $\Lambda_0 = \Lambda$. E.g. we know that the deSitter universe is a solution with $\Lambda > 0$, $\varepsilon = 0$ and $p = 0$. We may re-interpret it as a solution with $\Lambda = 0$, a constant density $\varepsilon > 0$ and a constant pressure $p = -\varepsilon < 0$, see Worksheet 4, Problem 2.

If we make special assumptions on ε and p , e.g. if we assume that p is related to ε by an equation of state, then (186) and (187) become a system of equations for $a(t)$ and $\varepsilon(t)$. Before solving this system for special cases, we consider the static solutions, i.e., we consider the Friedmann equations (186) and (187) for the case that $a(t) = a_0 = \text{constant}$:

$$\frac{3k}{a_0^2} - \Lambda = \kappa \varepsilon, \quad (189)$$

$$-\frac{k}{a_0^2} + \Lambda = \kappa p. \quad (190)$$

We have seen before that in the dust case ($p = 0$ and $\varepsilon = c^2 \mu \geq 0$) static solutions exist only for the cases $k = 1$ and $k = 0$. Now, with a pressure, we also have solutions with $k = -1$; however, as

$$\kappa(\varepsilon + p) = \frac{2k}{a_0^2}, \quad (191)$$

we see that in this case the pressure (or the density) has to be negative. We summarise the static solutions:

- $k = 1$: This is Einstein's static universe,

$$g = -c^2 dt^2 + a_0^2 \left(d\chi^2 + \sin^2 \chi (d\vartheta^2 + \sin^2 \vartheta^2 d\varphi^2) \right). \quad (192)$$

It is a dust solution with a positive cosmological constant ($\Lambda = a_0^{-2} > 0$, $\varepsilon = 2\kappa^{-1}a_0^{-2} > 0$ and $p = 0$), recall (164). However, by (188) we are free to re-interpret it, e.g., as a solution without a cosmological constant but with a positive density and a negative pressure.

- $k = 0$: This is Minkowski spacetime with the natural slicing associated with inertial systems,

$$g = -c^2 dt^2 + a_0^2 \left(d\chi^2 + \chi^2 (d\vartheta^2 + \sin^2 \vartheta^2 d\varphi^2) \right). \quad (193)$$

It is a vacuum solution without a cosmological constant ($\Lambda = 0$, $\varepsilon = 0$ and $p = 0$). Note, however, that it is possible to re-interpret it as a solution with a cosmological constant and a funny matter content.

- $k = -1$: This spacetime is known as static hyperbolic spacetime,

$$g = -c^2 dt^2 + a_0^2 \left(d\chi^2 + \sinh^2 \chi (d\vartheta^2 + \sin^2 \vartheta^2 d\varphi^2) \right). \quad (194)$$

We are free to choose the cosmological constant at will. Then ε and p are determined by (189) and (190),

$$\Lambda = \Lambda_0, \quad \varepsilon = -\frac{3}{\kappa a_0^2} - \frac{\Lambda_0}{\kappa}, \quad p = \frac{1}{\kappa a_0^2} + \frac{\Lambda_0}{\kappa}. \quad (195)$$

If we want to have ε positive, we have to choose $\Lambda_0 < -3/(\kappa a_0^2)$. But then p is necessarily negative.

Having the static solutions out of the way, we try to reduce the Friedmann equations to one first-order equation, as we did for a dust, by utilising a conservation law. From (186) we find that

$$\begin{aligned}\frac{d}{dt}\left(\frac{\kappa}{3}\varepsilon c^2 a^3\right) &= \frac{d}{dt}\left(c^2 k a + \left(\frac{da}{dt}\right)^2 a - \frac{\Lambda}{3}c^2 a^3\right) \\ &= c^2 k \frac{da}{dt} + \left(\frac{da}{dt}\right)^3 + 2 \frac{da}{dt} \frac{d^2 a}{dt^2} a - \Lambda c^2 a^2 \frac{da}{dt} \\ &= \frac{da}{dt} \left(c^2 k + \left(\frac{da}{dt}\right)^2 + 2 \frac{d^2 a}{dt^2} a - \Lambda c^2 a^2\right).\end{aligned}$$

By (187), this can be rewritten as

$$\frac{d}{dt}\left(\frac{\kappa}{3}\varepsilon c^2 a^3\right) = -c^2 \kappa p a^2 \frac{da}{dt},$$

hence

$$\frac{d}{dt}(\varepsilon a^3) = -p \frac{da}{dt} a^3. \quad (196)$$

This is the first law of thermodynamics for a volume element,

$$dU = T dS - p dV, \quad (197)$$

for the case of an isentropic process, $dS = 0$. Isentropic means that there is no heat transfer between the volume element and its neighbourhood; this assumption is implicitly included by requiring that the energy-momentum tensor has the form of a perfect fluid. Although we have energy conservation in the sense that no energy is produced, $\nabla_\mu T^{\mu\nu} = 0$, the energy in a comoving volume is *not* preserved because the pressure is doing work.

As long as ε and p are unrelated, the energy balance law (196) cannot be further specified. In particular, we cannot rewrite (196) in the form $d(\dots)/dt = 0$. However, if we assume an equation of state, then this is possible. We consider here only the special kind of an equation of state where the pressure is directly proportional to the energy density,

$$p(t) = w \varepsilon(t), \quad w = \text{constant}. \quad (198)$$

The energy-momentum tensor is then of the form

$$T_{\rho\sigma} = (\varepsilon + p) \frac{U_\rho U_\sigma}{c^2} + p g_{\rho\sigma} = \varepsilon \left((1 + w) \frac{U_\rho U_\sigma}{c^2} + w g_{\rho\sigma} \right). \quad (199)$$

The energy balance law (196) specifies to

$$\begin{aligned}\frac{d}{dt}(\varepsilon a^3) &= -w \varepsilon \frac{da}{dt} a^3, \\ \frac{d\varepsilon}{dt} a^3 + \varepsilon \frac{da^3}{dt} &= -w \varepsilon \frac{da^3}{dt} \\ \frac{d\varepsilon}{\varepsilon} + (1 + w) \frac{da^3}{a^3} &= 0,\end{aligned}$$

$$\ln(\varepsilon) + (1 + w) \ln(a^3) = C_0$$

with an integration constant C_0 , hence

$$\varepsilon(t) a(t)^{3(1+w)} = \text{constant}. \quad (200)$$

Three cases are of particular interest in view of applications:

(i) $w = 0$: This is the dust case we have already considered,

$$p = 0, \quad \varepsilon = \mu c^2, \quad (201)$$

$$T_{\rho\sigma} = \mu U_\rho U_\sigma. \quad (202)$$

The conservation law (200) reads

$$\mu(t) a(t)^3 = \text{constant} \quad (203)$$

which is just the statement that the mass in a comoving volume is constant. We know already that this is true for a dust, recall (166).

(ii) $w = -1$: This is a perfect fluid mimicking a cosmological constant,

$$p = -\varepsilon, \quad (204)$$

$$T_{\rho\sigma} = -\varepsilon g_{\rho\sigma}. \quad (205)$$

The conservation law (200) requires the energy density to be constant,

$$\varepsilon(t) = \text{constant}, \quad (206)$$

so the energy-momentum tensor has, indeed, the form of a cosmological term if we shift it to the left-hand side of the field equation, cf. Worksheet 4, Problem 2.

(iii) $w = 1/3$: This is an important case we have not yet treated so far. It describes a perfect fluid that models radiation in terms of a “photon gas”:

$$p = \frac{1}{3} \varepsilon, \quad (207)$$

$$T_{\rho\sigma} = \frac{4\varepsilon}{3c^2} U_\rho U_\sigma + \frac{\varepsilon}{3} g_{\rho\sigma}. \quad (208)$$

The trace of the energy-momentum tensor vanishes,

$$T_\rho{}^\rho = -\frac{4\varepsilon}{3c^2} c^2 + \frac{\varepsilon}{3} 4 = 0. \quad (209)$$

The conservation law (200) requires

$$\varepsilon(t) a(t)^4 = \text{constant}. \quad (210)$$

A rigorous justification of the statement that such a perfect fluid describes radiation would require a derivation from kinetic theory. We cannot do this here, but we will give two arguments indicating that the statement is true. The first argument

builds upon the fact that the trace of the energy-momentum tensor vanishes: The trace T_ρ^ρ of the energy-momentum tensor is a scalar, invariant under coordinate transformations. From the field equation we read that κT_ρ^ρ has the dimension $1/\text{length}^2$, so an energy-momentum tensor with non-vanishing trace defines a length scale. For a gas that consists of particles with a certain non-zero rest-mass m , this length scale is determined by the Schwarzschild radius associated with m . If the rest mass of the particles is zero, as it is for photons, such a length scale does not exist which means that the energy-momentum tensor must be trace-free. The second argument is based on the conservation law. For a dust that consists of massive particles, the rest mass of each particle remains constant, so the energy density $\varepsilon = c^2 \mu$ falls off with the volume, $\varepsilon \sim a^{-3}$. For a photon, the energy changes according to the redshift law, $1+z = a(t_o)/a(t_e)$, so the energy density gets a fourth factor of $1/a$ such that $\varepsilon \sim a^{-4}$ in accordance with (210).

With the conservation law (200) we may rewrite the first Friedmann equation (186) as

$$k a(t)^{1+3w} + \frac{1}{c^2} a'(t)^2 a(t)^{1+3w} - \frac{\Lambda}{3} a(t)^{3(1+w)} = \frac{\kappa}{3} \varepsilon(t) a(t)^{3(1+w)} =: a_0^{1+3w} \quad (211)$$

where the constant a_0 has the dimension of a length as can be read from comparing with the left-hand side. As for the dust case, we do not have to consider eq. (187) separately because it is automatically satisfied as long as $a' \neq 0$. We have now three parameters at our disposal: The discrete parameter k which takes the value $-1, 0$ or 1 , and the two parameters Λ and w which may take any real values. Here we will consider only the special case that

$$k = 0, \quad \Lambda = 0 \quad (212)$$

to concentrate on the influence of the pressure. Then the Friedmann equation (211) simplifies to

$$\begin{aligned} \frac{1}{c^2} a'(t)^2 a(t)^{1+3w} &= a_0^{1+3w}, \\ a^{(1+3w)/2} \frac{da}{dt} &= \pm c a_0^{(1+3w)/2}. \end{aligned} \quad (213)$$

We will solve this differential equation for the three cases which are of particular interest in view of applications.

- (i) $w = 0$: Just as a cross-check, we will re-examine the dust case. Then the differential equation (213) reads

$$\begin{aligned} a^{1/2} da &= \pm c a_0^{1/2} dt, \\ \frac{2}{3} a^{3/2} &= \pm c a_0^{1/2} (t - t_i) \end{aligned}$$

with an integration constant t_i . If we want to have a universe with a big bang at $t = 0$, we have to choose $t_i = 0$ and the plus sign,

$$a(t) = \left(\frac{9 a_0 c^2}{4} \right)^{1/3} t^{2/3} = b t^{2/3}.$$

This gives the Einstein-deSitter universe, as we had known before,

$$g = -c^2 dt^2 + b^2 t^{4/3} \left(d\chi^2 + \chi^2 (d\vartheta^2 + \sin^2 \vartheta d\varphi^2) \right).$$

- (ii) $w = -1$: This is a perfect fluid mimicking a cosmological constant. The differential equation (213) reads

$$a^{-1} da = \pm c a_0^{-1} dt ,$$

$$\ln(a) = \pm c a_0^{-1} t - C .$$

If we choose $C = \ln(a_0)$ and the plus sign we get

$$a(t) = a_0 \exp\left(\frac{ct}{a_0}\right)$$

which is, indeed, the deSitter universe with a flat slicing,

$$g = -c^2 dt^2 + a_0^2 \exp\left(\frac{2ct}{a_0}\right) \left(d\chi^2 + \chi^2 (d\vartheta^2 + \sin^2\vartheta d\varphi^2)\right) .$$

- (iii) $w = 1/3$: For a universe filled with radiation, the differential equation (213) reads

$$a da = \pm c a_0 dt ,$$

$$\frac{1}{2} a^2 = \pm c a_0 (t - t_i) .$$

If we want to have a universe with a big bang at $t = 0$ we have to choose $t_i = 0$ and the plus sign,

$$a(t) = \sqrt{2 a_0 c t} .$$

So in a radiation-filled universe the scale factor grows with $t^{1/2}$, in contrast to a dust-filled universe where it grows with $t^{2/3}$, see Figure 28.

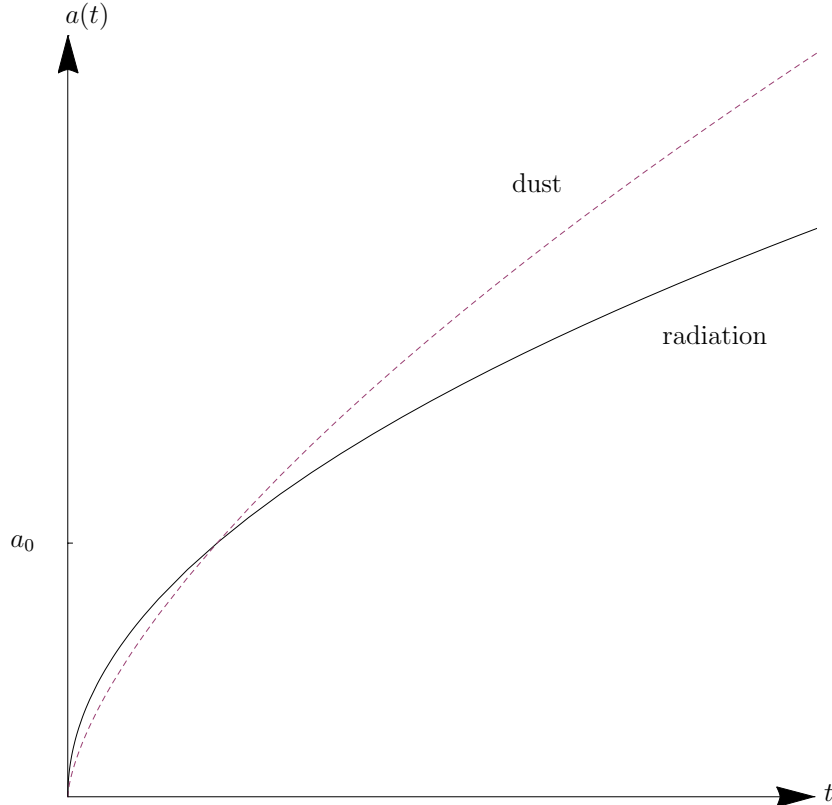


Figure 28: Radiation versus dust

Anticipating later discussions, we may assume as a rather realistic model of our universe a total energy-momentum tensor that is composed of two perfect fluids, one for a dust (modelling ordinary matter and also “cold” dark matter) $T_{\rho\sigma}^m$, and one for a photon gas (modelling the cosmic background radiation) $T_{\rho\sigma}^r$. If we also allow for the cosmological constant (as the simplest way of modelling dark energy), the field equation reads

$$R_{\rho\sigma} - \frac{R}{2} g_{\rho\sigma} + \Lambda g_{\rho\sigma} = \kappa (T_{\rho\sigma}^m + T_{\rho\sigma}^r). \quad (214)$$

We assume that the dust and the radiation are both at rest with respect to the standard observers of our Robertson-Walker spacetime,

$$T_{\rho\sigma}^m = \frac{\varepsilon_m}{c^2} U_\rho U_\sigma, \quad T_{\rho\sigma}^r = \frac{4\varepsilon_r}{3c^2} U_\rho U_\sigma + \frac{\varepsilon_r}{3} g_{\rho\sigma} \quad (215)$$

with $U^\rho = \delta_t^\rho$. Here $\varepsilon_m = c^2 \mu_m$ denotes the energy density of the matter and ε_r denotes the energy density of the radiation. Then the first Friedmann equation (119) reads

$$\frac{3c^2 k}{a(t)^2} + \frac{3}{a(t)^2} a'(t)^2 - \Lambda c^2 = \kappa c^2 (\varepsilon_m(t) + \varepsilon_r(t)), \quad (216)$$

hence

$$\varepsilon_m(t) + \varepsilon_r(t) + \frac{\Lambda}{\kappa} - \frac{3a'(t)^2}{\kappa c^2 a(t)^2} = \frac{3k}{\kappa a(t)^2}. \quad (217)$$

Evaluating this equation at a time t_o (“now”) yields

$$\varepsilon_m(t_o) + \varepsilon_r(t_o) + \frac{\Lambda}{\kappa} - \frac{3H(t_o)^2}{\kappa c^2} = \frac{3k}{\kappa a(t_o)^2}. \quad (218)$$

If we introduce, again, the critical density (177)

$$\varepsilon_c(t_o) = c^2 \mu_c(t_o) = \frac{3H(t_o)^2}{\kappa c^2}, \quad (219)$$

we can rewrite this equation as

$$\frac{\varepsilon_m(t_o)}{\varepsilon_c(t_o)} + \frac{\varepsilon_r(t_o)}{\varepsilon_c(t_o)} + \frac{\Lambda}{\kappa \varepsilon_c(t_o)} - 1 = \frac{3k}{\kappa a(t_o)^2 \varepsilon_c(t_o)}. \quad (220)$$

With the density parameters for dust (matter), radiation and cosmological constant,

$$\Omega_m(t_o) = \frac{\varepsilon_m(t_o)}{\varepsilon_c(t_o)}, \quad \Omega_r(t_o) = \frac{\varepsilon_r(t_o)}{\varepsilon_c(t_o)}, \quad \Omega_\Lambda(t_o) = \frac{\Lambda}{\kappa c^4 \varepsilon_c(t_o)}, \quad (221)$$

we get the famous relation

$$\Omega_m(t_o) + \Omega_r(t_o) + \Omega_\Lambda(t_o) \begin{cases} < 1 & \text{if } k = -1, \\ = 1 & \text{if } k = 0, \\ > 1 & \text{if } k = 1. \end{cases} \quad (222)$$

We have seen that in a universe filled with radiation alone the density falls off as a^{-4} , while in a universe filled with a dust alone it falls off as a^{-3} . This indicates that in

an expanding universe the radiation is important for the early universe but that its contribution becomes negligible for later times. We believe that at the present stage of our universe the radiation can be neglected. We also have good indications that our universe is spatially flat, $k = 0$. Then $\Omega_r(t_o)$ is negligibly small if t_o means “now”, and $\Omega_m(t_o) + \Omega_\Lambda(t_o) = 1$. Referring to the fact that in the present stage the dynamics of our universe is modelled by a cosmological constant and cold (pressureless) matter which includes dark matter, one speaks of the Λ CDM (Lambda-cold-dark-matter) model. We will come back to the observational foundations of this model later.

(d) Solutions with a scalar field as source

We have seen that, whenever a Robertson-Walker metric is plugged into Einstein’s field equation, the energy-momentum tensor on the right-hand side has the perfect-fluid form,

$$T_{\rho\sigma} = (\varepsilon + p) \frac{U_\rho U_\sigma}{c^2} + p g_{\rho\sigma}.$$

However, we may interpret this energy-momentum tensor in a different way. In this section we discuss the question of whether it can be interpreted as the energy-momentum tensor of a scalar field. This has important applications in cosmology. Several hypothetical scalar fields (or hypothetical “particles” in the quantised version) are discussed which may have a strong influence on the dynamics of the universe. The three most important of them are:

The Higgs field: In the basic version of gauge theories, all fields are massless. The Higgs field was invented to allow for massive fields. The mass terms come about by the interaction with the Higgs field. In 2012 a particle was detected at the Large Hadron Collider of CERN that is believed to be the Higgs particle. This won Peter Higgs and François Englert the Nobel Prize in Physics 2013. The Higgs field is a complex-valued scalar field that might have played an important role in cosmology at an early stage.

The inflaton: The inflaton field is a scalar field, in most theories assumed to be real-valued, that drives inflation. The idea is that it acted, for a period at a very early stage of the universe (something like 10^{-36} to 10^{-33} seconds after the big bang) like an enormously big cosmological constant, producing an exponential growth of the scale factor. The mechanism must be tuned in a way that the action of the inflaton field was then switched off so that it played no role for the later development of the universe. This is often called the “graceful exit” of inflation. The motivation for introducing an inflationary phase was that this would explain several things:

- The horizon problem: How could the universe become homogeneous at a time when, because of the existence of particle horizons, its different parts had had no time to interact?
- The monopole problem: Why is it that we do not observe a large number of magnetic monopoles although they are thought to have come into existence during phase transitions in the early universe?
- The flatness problem: Is it not highly unlikely that we live in a universe with spatial curvature K very close to zero if we think that our universe started with a random initial condition?

If there was an inflationary phase, different parts of the universe would have come into causal contact much earlier, the magnetic monopoles would have been diluted, and the blowing-up of the universe would bring the curvature K to a very small value whatever the initial conditions have been. Inflation was invented in 1980 independently by A. Starobinsky and by A. Guth. It was further developed by A. Linde, A. Albrecht, P. Steinhardt and many others. To date there exists a large variety of inflationary scenarios. The basic idea of inflation will be explained below in terms of the simplest of these scenarios (“slow-roll inflation”).

The quintessence: Since the late 1990s we have strong evidence that the expansion of our universe is accelerated. The easiest way to explain such an accelerated expansion is by introducing a positive cosmological constant. As we know, this may be re-interpreted as a perfect fluid with the equation of state $p = -\varepsilon$ (“dark energy”). We will see below that another possibility is to re-interpret the cosmological term as being produced by a (real-valued) scalar field. If we adopt this re-interpretation of the cosmological constant, we call it the quintessence field. (This name was introduced by L. Krauss around 2000. C. Wetterich, who was one of the pioneers of this idea, had called it the cosmon field.) In comparison to the inflaton, the quintessence field produces an exponential growth that is much smaller and becomes relevant at a much later time.

Several other hypothetical scalar fields (“phantom fields”, “chameleon fields”, “galileon fields” ...) have been suggested, but it seems fair to say that their existence is highly speculative. The three fields mentioned above provide the main motivation for us to study scalar fields and their coupling to gravity on a Robertson-Walker spacetime.

To that end we begin with the simplest type of a scalar field equation. We only consider real-valued fields ϕ . On Minkowski spacetime, the Klein-Gordon equation

$$\square\phi - m^2\phi = 0, \quad \square\phi = \eta^{\mu\nu}\partial_\mu\partial_\nu\phi = \Delta\phi - \frac{1}{c^2}\partial_t^2\phi \quad (223)$$

is the unique Lorentz-invariant linear field equation of second order for a scalar field. Here m is a constant that is related to the mass M associated with the field by $m = M/(c\hbar)$. (To put this another way, $1/m$ is the Compton wave-length of the field.) By the rule of minimal coupling, this generalises on a curved spacetime to the equation

$$\nabla^\mu\nabla_\mu\phi - m^2\phi = 0, \quad \nabla^\mu\nabla_\mu\phi = g^{\mu\nu}\nabla_\mu\nabla_\nu\phi = g^{\mu\nu}\left(\partial_\mu\partial_\nu\phi - \Gamma^\rho_{\mu\nu}\partial_\rho\phi\right). \quad (224)$$

In the following we consider a generalised Klein-Gordon equation,

$$\nabla^\mu\nabla_\mu\phi - V'(\phi) = 0, \quad (225)$$

with a potential $V(\phi)$. For the time being, V is largely arbitrary. We will only assume that it is bounded below, $V(\phi) \geq V_{\min}$, and that there is a ϕ_v with $V(\phi_v) = V_{\min}$. This “ground state” or “vacuum state” ϕ_v need not be unique. As the differential equation (225) is unchanged if we add a constant to V , we may assume without loss of generality that $V_{\min} = 0$, see Figure 29.

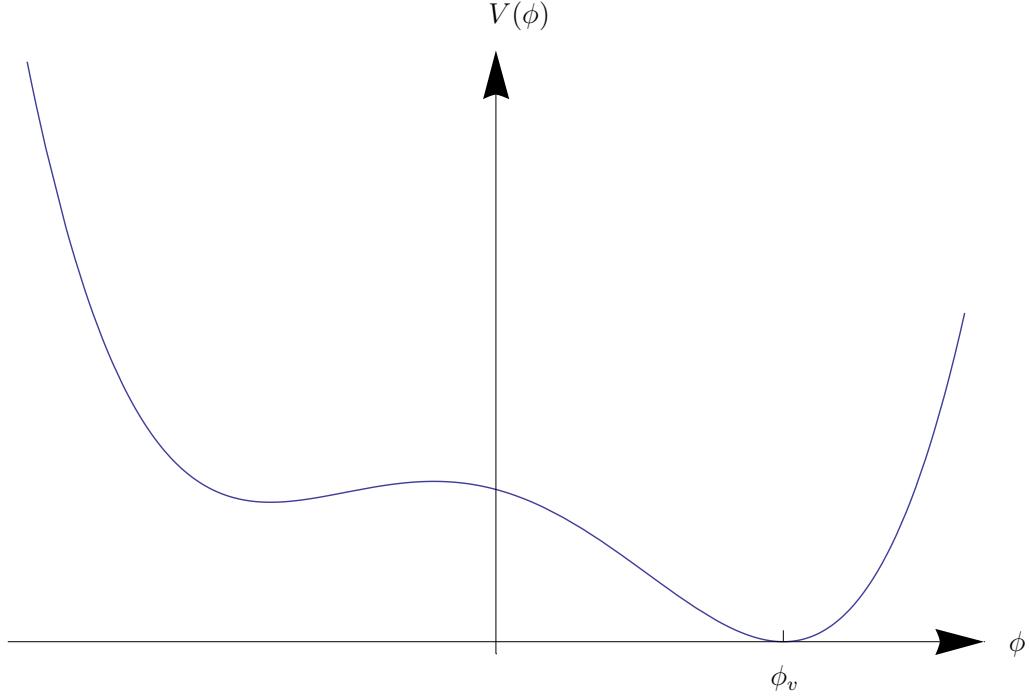


Figure 29: Potential $V(\phi)$

With the scalar field we associate the energy-momentum tensor

$$T_{\rho\sigma} = \nabla_\rho \phi \nabla_\sigma \phi - \left(\frac{1}{2} g^{\mu\nu} \nabla_\mu \phi \nabla_\nu \phi + V(\phi) \right) g_{\rho\sigma}. \quad (226)$$

We justify this by demonstrating that this energy-momentum tensor satisfies the energy conservation law $\nabla^\rho T_{\rho\sigma} = 0$ if ϕ solves the generalised Klein-Gordon equation and that the converse is also true unless $\nabla_a \phi = 0$. The covariant divergence of (226) give

$$\begin{aligned} \nabla^\rho T_{\rho\sigma} &= \nabla^\rho \left(\nabla_\rho \phi \nabla_\sigma \phi - \left(\frac{1}{2} g^{\mu\nu} \nabla_\mu \phi \nabla_\nu \phi + V(\phi) \right) g_{\rho\sigma} \right) \\ &= \nabla^\rho \nabla_\rho \phi \nabla_\sigma \phi + \nabla_\rho \phi \nabla^\rho \nabla_\sigma \phi - \left(\frac{1}{2} g^{\mu\nu} \left\{ \nabla^\rho \nabla_\mu \phi \nabla_\nu \phi + \nabla_\mu \phi \nabla^\rho \nabla_\nu \phi \right\} + V'(\phi) \nabla^\rho \phi \right) g_{\rho\sigma} \\ &= \nabla^\rho \nabla_\rho \phi \nabla_\sigma \phi + \nabla_\rho \phi \nabla^\rho \nabla_\sigma \phi - \frac{1}{2} g^{\mu\nu} \left\{ \nabla_\sigma \nabla_\mu \phi \nabla_\nu \phi + \nabla_\mu \phi \nabla_\sigma \nabla_\nu \phi \right\} - V'(\phi) \nabla_\sigma \phi \\ &= \nabla^\rho \nabla_\rho \phi \nabla_\sigma \phi + \nabla^\mu \phi \nabla_\mu \nabla_\sigma \phi - \nabla^\mu \phi \nabla_\sigma \nabla_\mu \phi - V'(\phi) \nabla_\sigma \phi. \end{aligned} \quad (227)$$

As covariant derivatives commute if they are applied to a scalar (!) field, the middle terms cancel, $\nabla_\mu \nabla_\sigma \phi = \nabla_\sigma \nabla_\mu \phi$, so

$$\nabla^\rho T_{\rho\sigma} = \nabla_\sigma \phi \left(\nabla^\rho \nabla_\rho \phi - V'(\phi) \right). \quad (228)$$

So we see that the generalised Klein-Gordon equation (225) implies the conservation law $\nabla^\rho T_{\rho\sigma} = 0$. Recall that the latter is always true for the energy-momentum tensor on the right-hand side of Einstein's field equation. It is, thus, consistent with Einstein's field equation to associate the energy-momentum tensor (226) with a scalar field that satisfies (225). Moreover, we read from (228) that conversely the energy-conservation law implies the generalised Klein-Gordon equation unless $\nabla^\sigma \phi = 0$.

We will now investigate if the energy-momentum tensor of a scalar field is of the same form as that for a perfect fluid. As we know that for a Robertson-Walker spacetime the energy-momentum tensor *always* has the form of a perfect fluid, this is a necessary step if we want to consider scalar fields on a Robertson-Walker spacetime.

We have to identify the two expressions

$$T_{\rho\sigma} = \nabla_\rho \phi \nabla_\sigma \phi - \left(\frac{1}{2} g^{\mu\nu} \nabla_\mu \phi \nabla_\nu \phi + V(\phi) \right) g_{\rho\sigma} \quad (229)$$

and

$$T_{\rho\sigma} = (\varepsilon + p) \frac{U_\rho U_\sigma}{c^2} + p g_{\rho\sigma}. \quad (230)$$

By transvecting with U^σ one finds that this is possible only if $\nabla_\rho \phi$ and U_ρ are linearly dependent,

$$\nabla_\rho \phi = s U_\rho \quad (231)$$

with a scalar factor s . This factor is determined by the normalisation condition on the four-velocity, $U^\rho U_\rho = -c^2$. We find

$$\nabla^\rho \phi \nabla_\rho \phi = -c^2 s^2. \quad (232)$$

So the identification requires that either $\nabla^\rho \phi$ is timelike (and $s \neq 0$) or $\nabla^\rho \phi = 0$ (and $s = 0$).

We can now determine p and ε in terms of the scalar field by equating the two expressions of the energy-momentum tensor. Comparing the coefficients in front of $g_{\rho\sigma}$ yields

$$p = -\frac{1}{2} \nabla^\mu \phi \nabla_\mu \phi - V(\phi). \quad (233)$$

Equating the first terms requires

$$(\varepsilon + p) \frac{U_\rho U_\sigma}{c^2} = s^2 U_\rho U_\sigma, \quad \varepsilon + p = -\nabla^\mu \phi \nabla_\mu \phi, \quad (234)$$

$$\varepsilon = -\frac{1}{2} \nabla^\mu \phi \nabla_\mu \phi + V(\phi). \quad (235)$$

We summarise our results in the following way: The energy-momentum tensor of a scalar field is of the form of a perfect fluid whenever the gradient of the scalar field is timelike or zero. The corresponding pressure and the corresponding energy density are then given by (233) and (235).

Now we specify to the case that we are on a Robertson-Walker spacetime,

$$g_{\mu\nu} dx^\mu dx^\nu = -c^2 dt^2 + a(t)^2 \left(d\chi^2 + \xi(\chi)^2 (d\vartheta^2 + \sin^2\vartheta d\varphi^2) \right). \quad (236)$$

If we want to consider a scalar field as the source term in Einstein's field equation for such a spatially homogeneous universe, the scalar field must be independent of the spatial coordinates for consistency,

$$\phi = \phi(t). \quad (237)$$

This guarantees that $\nabla_\rho \phi = \delta_\rho^t d\phi/dt$ is indeed timelike or zero. Then the equations for pressure and density, (233) and (235), simplify to

$$p = -\frac{1}{2} g^{tt} \left(\frac{d\phi}{dt} \right)^2 - V(\phi), \quad \varepsilon = -\frac{1}{2} g^{tt} \left(\frac{d\phi}{dt} \right)^2 + V(\phi). \quad (238)$$

With $g^{tt} = 1/g_{tt} = -1/c^2$ this results in

$$p = \frac{1}{2c^2} \left(\frac{d\phi}{dt} \right)^2 - V(\phi), \quad \varepsilon = \frac{1}{2c^2} \left(\frac{d\phi}{dt} \right)^2 + V(\phi). \quad (239)$$

Clearly, p and ε are not in general related by an equation of state. An equation of state results in two very special cases:

If $V(\phi) = 0$ we have $p = w\varepsilon$ with $w = 1$.

If $d\phi/dt = 0$ we have $p = w\varepsilon$ with $w = -1$.

Recall that $w = -1$ is the case of a perfect fluid mimicking a cosmological constant. As $V(\phi) \geq 0$, we have in any case

$$-1 \leq \frac{p}{\varepsilon} = \frac{\frac{1}{2c^2} \left(\frac{d\phi}{dt} \right)^2 - V(\phi)}{\frac{1}{2c^2} \left(\frac{d\phi}{dt} \right)^2 + V(\phi)} \leq 1. \quad (240)$$

With ε and p expressed in terms of the scalar field, the Friedmann equations read

$$\frac{3c^2 k}{a^2} + \frac{3}{a^2} \left(\frac{da}{dt} \right)^2 - \Lambda c^2 = \kappa c^2 \left(\frac{1}{2c^2} \left(\frac{d\phi}{dt} \right)^2 + V(\phi) \right), \quad (241)$$

$$-k - \frac{1}{c^2} \left(\frac{da}{dt} \right)^2 - \frac{2a}{c^2} \frac{d^2 a}{dt^2} + \Lambda a^2 = \kappa a^2 \left(\frac{1}{2c^2} \left(\frac{d\phi}{dt} \right)^2 - V(\phi) \right). \quad (242)$$

We know that Einstein's field equation implies $\nabla^\rho T_{\rho\sigma} = 0$ and that for a Robertson-Walker spacetime this equation takes the form of the energy balance equation (196). As

in the case of a scalar field the equation $\nabla^\rho T_{\rho\sigma} = 0$ is equivalent to the generalised Klein-Gordon equation (225), provided that $\nabla_\rho \phi \neq 0$, we can derive the special form of the equation (225) for a Robertson-Walker spacetime by evaluating the energy balance (196). This spares us the trouble of calculating the Christoffel symbols for the Robertson-Walker spacetime.

Eq. (196), which is just a version of the First Law of Thermodynamics, can be rewritten as

$$\begin{aligned}\frac{d\varepsilon}{dt} a^3 + \varepsilon 3 a^2 \frac{da}{dt} &= -p 3 a^2 \frac{da}{dt}, \\ \frac{d\varepsilon}{dt} &= -\frac{3}{a} (\varepsilon + p) \frac{da}{dt}.\end{aligned}\tag{243}$$

Inserting (233) and (235) yields

$$\begin{aligned}\frac{d}{dt} \left(\frac{1}{2 c^2} \left(\frac{d\phi}{dt} \right)^2 + V(\phi) \right) &= -\frac{3}{a c^2} \left(\frac{d\phi}{dt} \right)^2 \frac{da}{dt}, \\ \frac{1}{c^2} \frac{d\phi}{dt} \frac{d^2\phi}{dt^2} + V'(\phi) \frac{d\phi}{dt} &= -\frac{3}{a c^2} \left(\frac{d\phi}{dt} \right)^2 \frac{da}{dt}.\end{aligned}\tag{244}$$

For $d\phi/dt$ non-zero, this gives us the generalised Klein-Gordon equation on a Robertson-Walker spacetime,

$$\frac{1}{c^2} \frac{d^2\phi}{dt^2} + \frac{3}{c^2} \frac{1}{a} \frac{da}{dt} \frac{d\phi}{dt} + V'(\phi) = 0.\tag{245}$$

This equation is of the same form as the equation of motion for a particle in the one-dimensional potential V , with a damping term. The damping is proportional to the Hubble function $H(t) = a(t)^{-1} da(t)/dt$. (We assume that we are in an expanding universe, i.e., that $H(t)$ is positive.) So we may visualise the dynamics of the scalar field as the dynamics of a particle in the potential V with friction. This analogy is habitually used for scalar fields in cosmology, where one says that “the field is rolling down a slope of the potential” and that “it settles in a minimum after the oscillations are damped away”.

When considering scalar fields as the source in Einstein’s equation for a Robertson-Walker spacetime we have to solve the equations (241), (242) and (245). Actually, only two of these three equations are independent: (241) and (242) imply (245) unless $d\phi/dt = 0$ whereas (241) and (245) imply (242) unless $da/dt = 0$. If we exclude the static case $da/dt = 0$, we are left with two coupled ordinary first-order differential equations for the two unknown functions $a(t)$ and $\phi(t)$.

We will first demonstrate how a scalar field can act as a cosmological constant; we will then briefly discuss the “slow-roll inflation” scenario. For imitating a cosmological constant it is, of course, reasonable to consider the equations with $\Lambda = 0$. In addition, we restrict to the spatially flat case, so we assume

$$k = 0, \quad \Lambda = 0.\tag{246}$$

It is our goal to mimic the cosmological constant with a constant scalar field,

$$\phi(t) = \phi_0 = \text{constant} . \quad (247)$$

Then (245) requires

$$V'(\phi_0) = 0 , \quad (248)$$

i.e., the scalar field must sit in an extremum of the potential. This may be a minimum, a maximum or a saddle. If we want the solution to be stable with respect to small perturbations, we have to choose a local minimum, see Figure 30.

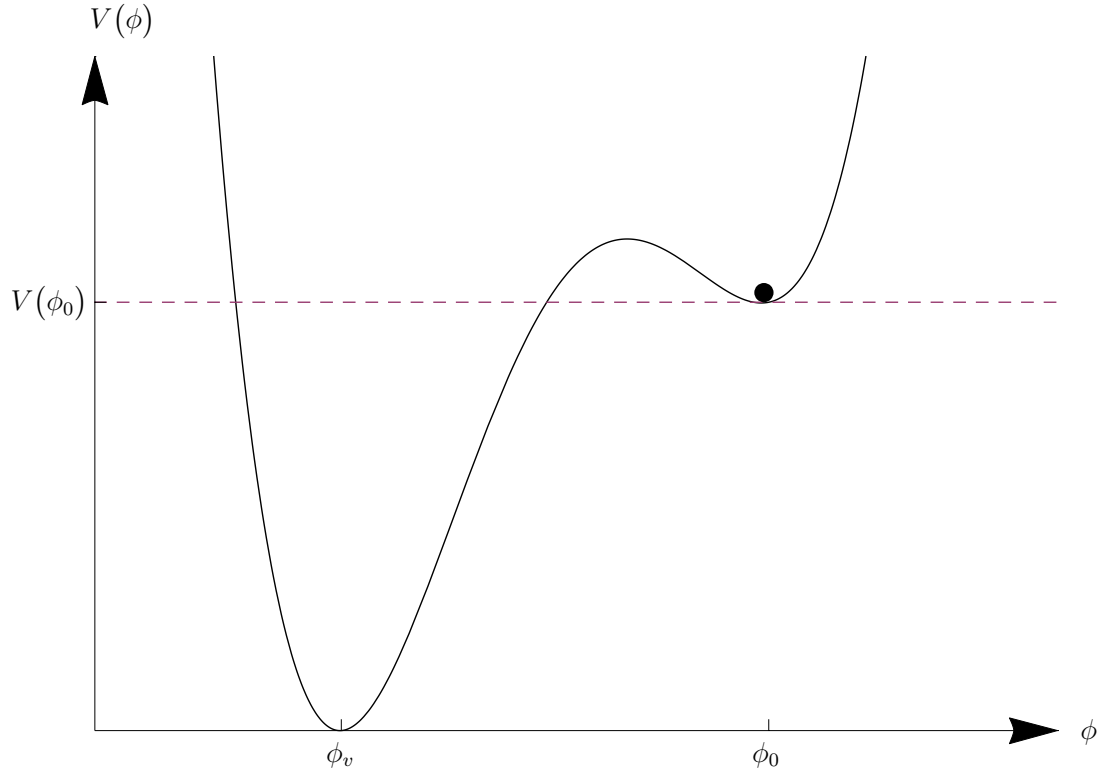


Figure 30: Potential $V(\phi)$ for quintessence

(241) reduces to

$$\begin{aligned} \frac{1}{a^2} \left(\frac{da}{dt} \right)^2 &= \frac{\kappa}{3} c^2 V(\phi_0) , \\ \frac{da}{a} &= \pm c \sqrt{\frac{\kappa}{3} V(\phi_0)} dt , \\ \ln(a) - \ln(a_0) &= \pm c \sqrt{\frac{\kappa}{3} V(\phi_0)} t , \\ a(t) &= a_0 \exp \left(\pm \sqrt{\frac{\kappa}{3} V(\phi_0)} c t \right) . \end{aligned} \quad (249)$$

With $\phi(t)$ and $a(t)$ determined by (247) and (249), respectively, (242) is satisfied as well.

If we have chosen the vacuum state, $\phi_0 = \phi_v$, we have $V(\phi_0) = 0$ and the scale factor is constant, $a(t) = a_0$. As we are considering the case $k = 0$, this gives us Minkowski spacetime. If, however,

$$V(\phi_0) > 0,$$

we may choose the solution with the plus sign and the integration constant as

$$a_0 = \sqrt{\frac{3}{\kappa V(\phi_0)}}.$$

Then we get deSitter spacetime with the flat slicing (i.e., half of the hyperboloid, recall Figure 16),

$$a(t) = \sqrt{\frac{3}{\kappa V(\phi_0)}} \exp\left(\sqrt{\frac{\kappa}{3} V(\phi_0)} c t\right). \quad (250)$$

The exponential growth of the scale factor is now driven by the scalar field. It has the same effect as a positive cosmological constant,

$$\Lambda \hat{=} \kappa V(\phi_0). \quad (251)$$

To explain the accelerated expansion of the universe *now* as indicated by observations of type Ia supernovae (see next chapter), we need a cosmological constant $\sqrt{\Lambda} \approx 10^{-26} \text{m}^{-1}$. We have already seen that this can be re-interpreted as the effect of a perfect fluid (“dark energy”) with the exotic equation of state $p = -\varepsilon$. We have now another possible interpretation: We may say that what seems to be a positive cosmological constant is actually a constant scalar field. This scalar field is called “quintessence”. Just as the re-interpretation as a perfect fluid, this makes the cosmological constant dynamical in the sense that it is not necessarily exactly a constant: The quintessence field may change in the course of time. If this is true, the future of our universe is not predictable; if there is a cosmological constant in the strict sense, it will dominate all other sources (radiation and matter) in the course of time so that the universe asymptotically approaches the deSitter spacetime.

The quintessence field is a possible re-interpretation of the cosmological constant but there is no compelling reason for introducing it. You may leave the cosmological term on the left-hand side of the field equation, or you may write it on the right-hand side and interpret it as a perfect fluid or as a scalar field. As long as all observations can be explained with a cosmological constant that is really a constant, this is largely a matter of taste. The situation is different with the inflaton: Just as the quintessence field, the inflaton field is supposed to create exponential growth, but at a much, much earlier stage of the universe and by a much, much bigger factor. Most importantly, the inflaton field should switch itself off at the end of the inflationary period. So here we are not re-interpreting a cosmological constant, we are rather considering a scalar field that acts *approximately* as a cosmological term only over a certain period of time. So we need a potential that is almost constant (i.e., very flat) over a certain ϕ interval, see Figure 31. While the scalar field is “slowly rolling” down this very flat slope, our

calculation for a constant scalar field ϕ_0 holds *approximately*, i.e., the scale factor grows *approximately* exponentially according to (250). If the scalar field rolls into the vacuum state ϕ_v , the exponential growth stops, because after a few oscillations the “damping term” in the generalised Klein-Gordon equation forces the scalar field to settle at ϕ_v which, as demonstrated above, leads to a constant scale factor. In this way, the inflaton switches itself off and plays no role in the later development of the universe. This is what one calls the “graceful exit” of inflation. By choosing $V(\phi_0)$ sufficiently big, the exponential growth rate during the inflationary period can be as large as we wish, see (250). We will roughly estimate in Worksheet 7 how big this growth rate should be in order to solve the horizon problem.

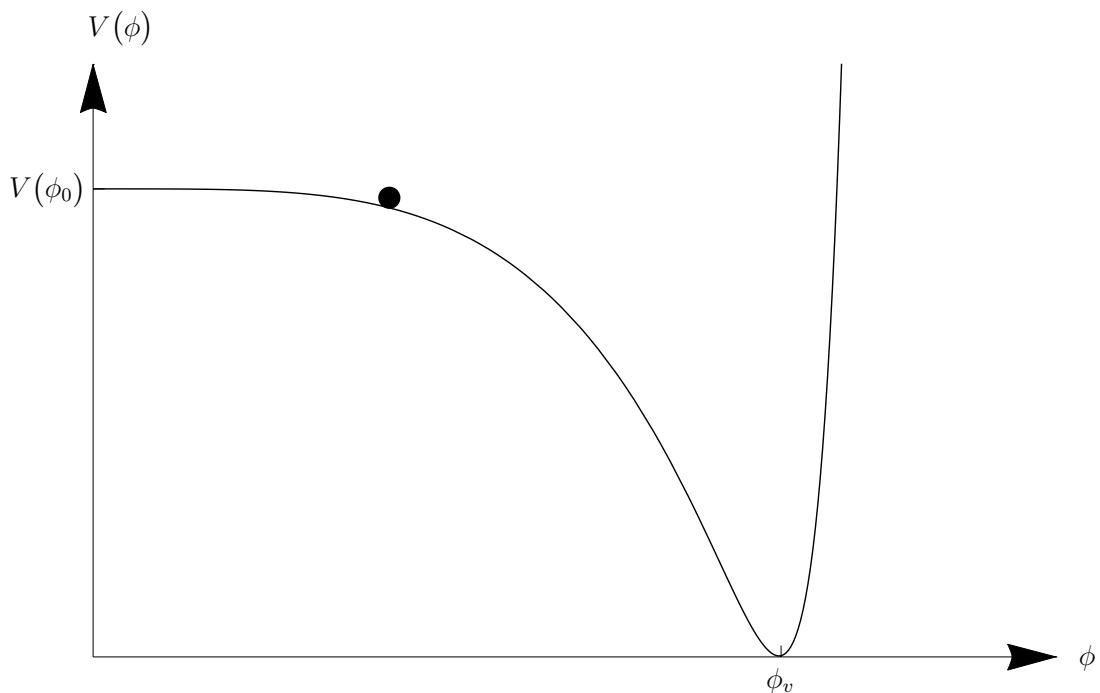


Figure 31: Potential $V(\phi)$ for the inflaton field

This so-called “slow-roll” inflationary scenario is the simplest way in which an inflationary stage can be produced. It was suggested by A. Linde and independently by A. Albrecht and P. Steinhardt in 1982. Already earlier, other inflationary scenarios had been introduced by A. Guth and by A. Starobinsky. Guth, who introduced the name “inflation”, wanted to identify the field that drives inflation with the Higgs field and he suggested a certain tunnelling process for the transition into the vacuum state. Starobinsky used a different approach, based on a modified field equation that was motivated by ideas from quantum gravity. In the mean-time there is a big number of different inflationary scenarios. By now, none of them is unanimously accepted. In any case, the presence of an inflationary period with a sufficiently big growing factor solves the horizon problem and the monopole problem. It is a matter of debate if it also solves the flatness problem

because in most inflationary models the fine-tuned initial conditions for the geometry are replaced by a fine-tuned initial condition for the scalar field. It should also be mentioned that inflation is still a hypothetical concept, not directly verified by observations. It is believed by a majority of cosmologists that an inflationary period took place at an early stage of the universe, but there are also outspoken critics, e.g. Roger Penrose.

We have now discussed the dynamics of Robertson-Walker universes for various matter sources. According to the model that is favoured by the majority of cosmologists (“concordance model”) we live in a Robertson-Walker universe with $k = 0$ where different matter sources were dominating at different stages; correspondingly, the scale factor changed with time in a different way at different stages. In the following we list these different stages; we will discuss in the next chapter on what observational evidence the concordance model is based and at what times the different periods took place.

- Big bang: We believe that the universe began with a hot big bang, a state of extreme density. The time immediately after the big bang is not yet theoretically understood. A (not yet existing) quantum theory of gravity is probably needed.
- Inflationary period: We conjecture that there was a very short period during which the scale factor grew exponentially by an enormous factor (10^{28} at least). Inflation was driven by a hypothetical scalar field called the inflaton. At the end of the inflationary period the inflaton field settled into the ground state and had no effect from that time on.
- Radiation-dominated period: After the graceful exit from inflation radiation gave the dominating contribution to the density. The density parameters of matter and of the cosmological constant were negligibly small in comparison to that of radiation. The universe was expanding $\sim t^{1/2}$, i.e., the expansion was decelerating.
- Λ CDM period: At present, the universe is matter-dominated and the effect of a positive cosmological constant has to be taken into account. The matter content can be modelled as a dust, i.e., as “cold matter”. It comprises the usual (“baryonic”) matter and a mysterious “dark matter” whose nature is unknown as of now. (CDM stands for “cold dark matter” which is short for “cold matter part of which is dark”.) The density parameter of radiation is now negligible; recall that the density of radiation falls off with a^{-4} while the density of matter (dust) falls off with a^{-3} . The cosmological constant may be re-interpreted as a perfect fluid with equation of state $p = -\varepsilon$ (“dark energy”) or as a scalar field (“quintessence”). Because of the cosmological constant, the expansion of the universe is accelerating.
- Asymptotic deSitter period: If the cosmological constant is really a constant, it will dominate the dynamics at late times. Our universe will then asymptotically approach the deSitter spacetime. If the cosmological constant is dynamical (i.e., a perfect fluid or a scalar field), it may change with time and the future of the universe cannot be predicted.

4 Observations

In this chapter we will summarize the observational facts on which the concordance model is based. The most important information comes from the cosmic background radiation and from the distance-redshift relation, but lensing and some other observational facts also provide important restrictions on the model.

4.1 Evidence for dark matter

There are several observational facts indicating that a large part of matter in the universe is dark and therefore detectable only by way of its gravitational field. As long as Einstein's general relativity theory (and, where applicable, the Newtonian approximation) is considered as the theoretical basis, these observational facts are compelling. We list them in chronological order.

- Velocity distribution in galaxy clusters: In 1933 and 1937 F. Zwicky wrote two papers in which he gave the first evidence for the existence of dark matter. He analysed the dynamics inside the Coma cluster, a galaxy cluster with more than 1000 galaxies. He found that the galaxies are too fast for the cluster to form a gravitationally bound system if one assumes that the total mass of the cluster is given by the luminous matter we are observing. He conjectured that more than 99.8 % of the matter is dark. (In the 1933 article, which is written in German, he speaks of “dunkle (kalte) Materie”.) We believe today that this number is too big. The reason is that at this time the distance ladder was wrongly calibrated: Zwicky assumed that the Coma cluster is about 15 Mpc away; actually it is more than 100 Mpc. If one re-analyses his calculation with a corrected distance scale, one finds that about 95 % of the matter in galaxy clusters must be dark. Zwicky's prediction of the existence of dark matter was largely ignored at the time.
- Rotation curves of galaxies: In the 1970s V. Rubin analysed rotation curves in a large sample of spiral galaxies, partly in collaboration with K. Ford. She looked at galaxies which are seen almost edge-on. Then, because of the rotation of the stars about the centre of the galaxy, the spectral lines are red-shifted on one side and blue-shifted on the other. Measuring these shifts in different parts of the galaxy gives the so-called rotation curve, i.e., the orbital velocity as a function of the radius. If all the mass of the galaxy were in the centre, the stars would move in a potential $\sim r^{-1}$ according to Newtonian gravity (which is a valid approximation here). This would give an orbital velocity $\sim 1/\sqrt{r}$. If one takes the actual distribution of the matter in the bulge and in the disc into account, according to the visible masses, one gets a curve that increases in the inner part, but then falls off like $1/\sqrt{r}$ in the outer part. This fall-off was not what V. Rubin observed: Actually, the rotation curves remained almost flat. This could be explained, on the basis of Einstein's general relativity and its Newtonian approximation, only by the assumption that the galaxy is embedded in a “dark matter halo”. V. Rubin estimated, that it should make up about 50 % of the galaxy's mass. Later observations indicated that it should be at least 85 %. Figure 32 shows the example of the galaxy NGC 3198. (The rotation curves of all galaxies show the same tendency.) The graph labelled “disk” is what one would see if there were only the visible matter. The graph labelled “halo” shows the

difference between this graph and the observed rotation curve which is interpreted as the effect of the dark matter halo.

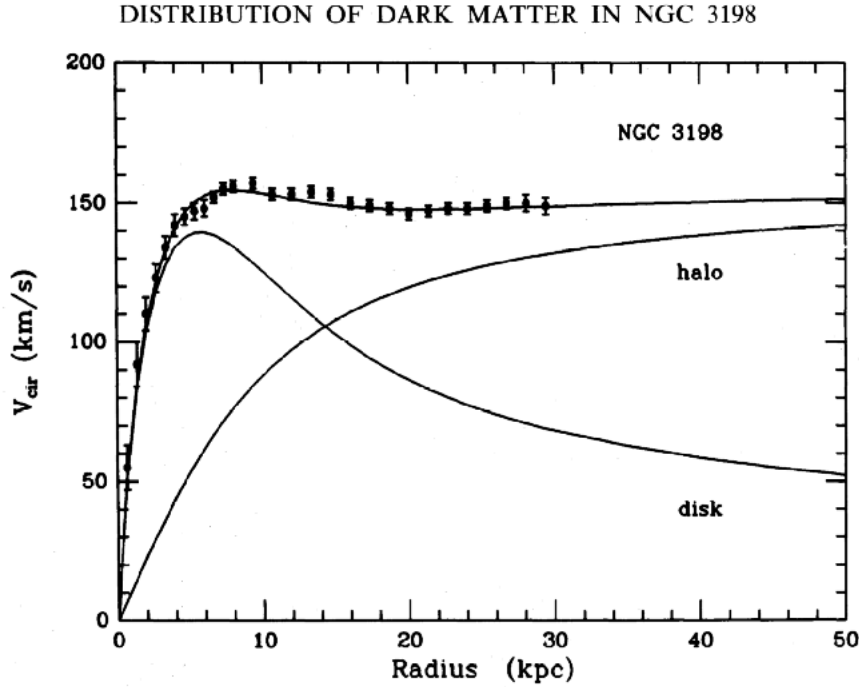


Figure 32: Rotation curve of the galaxy NGC 3198 [picture from T. S. van Albada, J. N. Bahcall, K. Begeman, R. Sancisi: *Astrophys. J.* 295, 305 (1985)]

The dark matter halo is usually modelled as spherically symmetric and several density profiles have been suggested, e.g.

Non-singular isothermal sphere: $\mu(r) = \frac{\mu_0 r_0^2}{r_0^2 + r^2}$,

Navarro-Frenk-White profile: $\mu(r) = \frac{\mu_0 r_0^4}{r^2 (r_0^2 + r^2)}$,

Einasto profile: $\mu(r) = \mu_0 \exp(-Ar^\alpha)$,

where μ_0 , r_0 , A and α are parameters that can be fitted to observations. The Navarro-Frenk-White profile is the most commonly used for the density of dark matter, not only in galaxies but also in galaxy clusters. It was found not just by guess-work but rather by numerical N-body simulations.

- Microlensing: The first candidates for dark matter in the halo of our galaxy that come to mind are “Massive Compact Halo Objects” (MaCHOS) such as black holes, brown dwarfs and planets. They cannot be seen directly, because they do not emit light, but they can be observed by the influence of their gravitational field on light: If a star passes behind such a compact object, the light is focussed towards the observer, i.e., one sees a light curve that goes up and then down again. The closer the star comes to the line of sight, the bigger is the effect, see Figure 33.

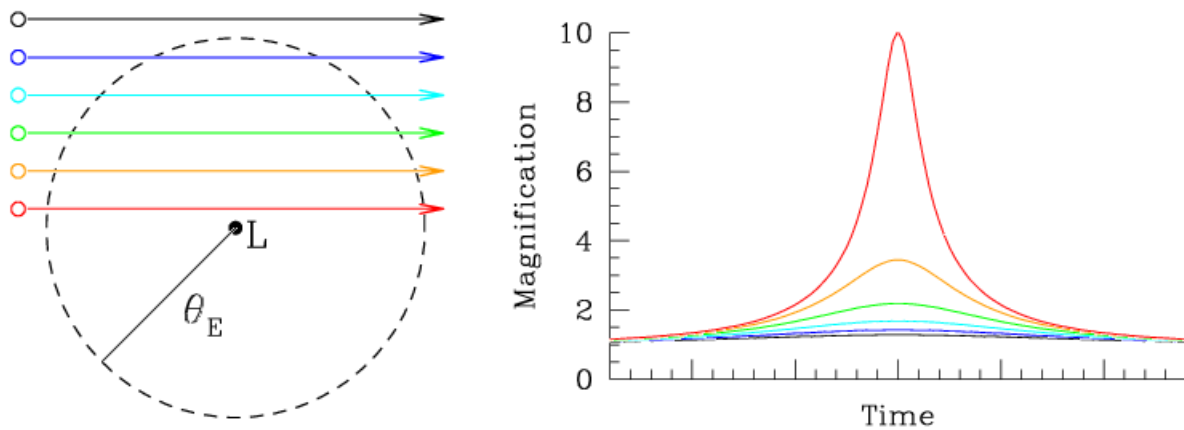


Figure 33: Microlensing

This is called microlensing. More precisely, the word microlensing is used for lensing situations where multiple images are created but cannot be resolved; so what one sees is just a change of brightness of the compound image. Microlensing events are routinely observed since the early 1990s. They are very common (several hundreds per year), and in the majority of cases the observations are made towards the halo of our galaxy. (Observations are also made towards the bulge of our galaxy, towards the Magellanic clouds and towards the Andromeda galaxy.) These observations give an upper bound on the total mass of the MaCHOs in our galaxy. The microlensing surveys came to the conclusion that not more than 20 % of the dark matter that is needed for explaining the rotation curves can be MaCHOs.

- Weak lensing: While microlensing is the most important tool for detecting dark matter in our galaxy or nearby, weak lensing is the most important tool for detecting dark matter in distant galaxy clusters. What one observes is the deformation of background galaxies by the lensing effect of the cluster. For understanding the basic idea, let us assume for a moment that the background galaxies were perfectly spherical. Then we would see each galaxy distorted into an ellipse on the sky, and the orientations and the eccentricities of the ellipses would tell us where the deflecting mass is located in the sky and how big its surface mass density (mass density projected onto the plane perpendicular to the view-line) is. Unfortunately, galaxies are not spherical. For most of them the shape can be approximated reasonably well by an ellipsoid. So when we see an ellipse in the sky we have to distinguish the intrinsic shape from the distortion effect produced by lensing. This can be done statistically, based on the assumption that the intrinsic shapes are distributed randomly. A sophisticated numerical method has been established to deduce from the observed shapes of background galaxies the surface mass density of a galaxy cluster. (The surface mass density is the volume mass density integrated over the line of sight, i.e., it has the dimension mass per area.) This has been worked out since the 1980s for a large number of galaxy clusters. If Einstein's theory describes the effect of gravitational fields on light correctly, all these observations confirm Zwicky's prediction that the majority of matter in galaxy clusters is dark.

The most famous example is the *bullet cluster*. These are two colliding galaxy clusters, see Figure 34. This picture consists of three different contributions which are overlayed: An ordinary photograph in the optical made with the Hubble space telescope (all the white or yellowish spots which are actually galaxies), an X-ray picture taken by the X-ray satellite Chandra (the two red clouds) and the surface mass density calculated from weak lensing observations (the two blue clouds).

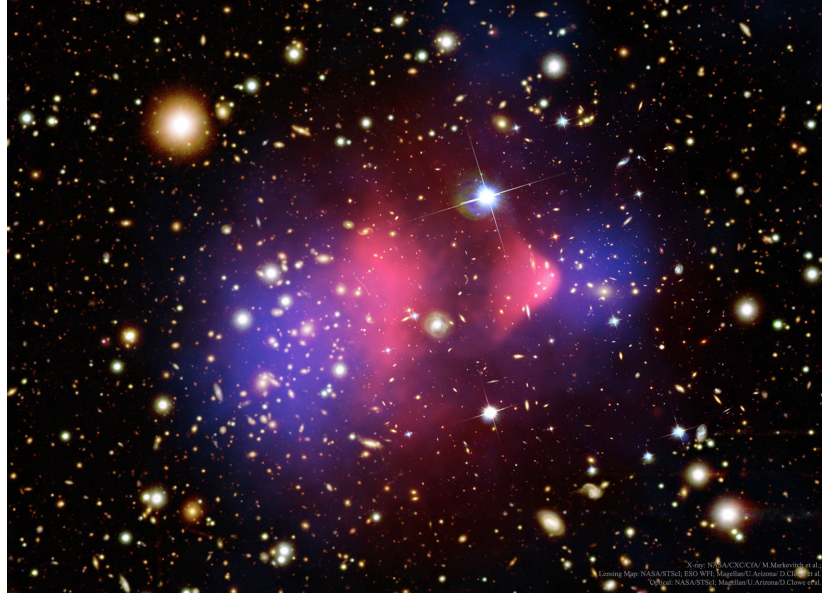


Figure 34: The bullet cluster [picture from apod.nasa.gov]

It is the shape of the red cloud on the right that gave the name to the bullet cluster, because it looks like a bullet rushing into a target. The interpretation is as follows: We see two colliding galaxy clusters. The stars in the galaxies move through each other largely without any effect, because collisions are rare. The hot gases, however, strongly decelerate each other when colliding, so they stay behind; this is what the red clouds are showing. As the majority of visible masses in a galaxy cluster is in the form of hot gases, one would expect the blue clouds to coincide with the red clouds. However, this is not what we see: The blue clouds have not been decelerated by the collision, so the majority of the gravitating mass must consist of a kind of dark matter that is more or less frictionless. The bullet cluster gives compelling evidence for the existence of dark matter in galaxy clusters, provided one accepts Einstein's general relativity theory, and it gives strong restrictions on the way this dark matter can interact with itself and with other matter. Since the discovery of the bullet cluster a few other pairs of colliding clusters with similar properties have been found.

Taking the evidence from the velocity distributions in galaxy clusters, from the rotation curves in galaxies and from weak lensing together, we are more or less forced to assume that about 90 % of the matter is dark. As the bullet cluster shows most clearly, the mysterious dark matter can interact only very weakly with other things and with itself.

Several hypothetical particles have been suggested as dark matter candidates, e.g.

- weakly interacting massive particles (WIMPs)
- axions,
- new types of neutrinos,
- ...

In spite of intensive searches, none of them has been detected so far. So the present status is: We do not know what this dark matter is, but we have to assume that it is there in order to explain the observations.

The only alternative to accepting the existence of dark matter seems to be a modification of the gravitational theory. As the observations that led us to postulating dark matter are mainly done at a level where the Newtonian approximation is valid, it would be necessary to modify already the Newtonian theory (and then Einstein's theory in a way that it gives the modified Newtonian theory in the appropriate limit). Several modified theories of this kind have been brought forward:

- Modified Newtonian Dynamics (MoND)

This theory was suggested by M. Milgrom in 1983. It modifies the Newtonian equation of motion (Newton's Lex Secunda) from $\vec{F} = m \vec{a}$ to

$$\vec{F} = m \mu\left(\frac{a}{a_0}\right) \vec{a}. \quad (252)$$

Here a_0 is a hypothetical constant of Nature with the dimension of an acceleration and μ is a function that is to be chosen in a way that the old version of the Lex Secunda is still valid if the acceleration a is much bigger than a_0 , i.e.,

$$\mu(x) \approx 1 \quad \text{for } x \gg 1. \quad (253)$$

Milgrom has demonstrated that the rotation curves of galaxies can be explained in a quite satisfactory manner if

$$a_0 \approx 10^{-10} \text{ m/s}^2 \quad (254)$$

and

$$\mu(x) = \frac{1}{1 - \frac{1}{x}} \quad (255)$$

or

$$\mu(x) = \frac{1}{\sqrt{1 - \frac{1}{x^2}}}. \quad (256)$$

Of course, MoND cannot be considered as anything else but a Newtonian-like limit of a "true theory" which generalises Einstein's theory.

- Tensor-Vector-Scalar (TeVeS) theory

It took about 20 years to find a generalisation of general relativity that reduces to MoND in the appropriate limit. It was found by J. Bekenstein in 2004. In this theory the gravitational field is not just described by a tensor field, as in Einstein's theory, but in addition by vector and scalar fields. The field equations are extremely complicated and the geometrical appeal of Einstein's theory is largely destroyed. TeVeS (and, thus, MoND) has problems to explain the observations of binary pulsars and of the cosmic background radiation. However, the greatest challenge for this theory is the bullet cluster. Milgrom admitted that he was not able to fully explain the observations of the bullet cluster within TeVeS/MoND.

- Conformal gravity

In 1989 P. Mannheim suggested that the rotation curves of galaxies can be explained in a theory where the gravitational field is still given by a metric tensor, as in general relativity, but the field equation is modified. Instead of Einstein's field equation, which derives from an action given by the Ricci scalar, this field equation derives from an action given by the square of the conformal curvature tensor (also known as the Weyl tensor). As a result, the left-hand side of the field equation is conformally invariant, i.e., it does not change if the metric is multiplied with a positive function. The same field equation was suggested already in 1920 by R. Bach. The theory suffers from severe conceptual problems. In particular, the conformal symmetry has to be broken by some mechanism in order to allow for non-zero masses because an energy-momentum tensor can be conformally invariant only in the case that it describes matter made up of massless particles (such as a photon gas).

4.2 The distance-redshift relation

Recall that we have found various versions of a “Hubble law” in Robertson-Walker spacetimes. Without using the field equation, we could demonstrate the following.

- There is an *exact* linear relation between proper distance D_p and proper velocity dD_p/dt , recall (92). This, however, is of no relevance in view of observations because D_p cannot be measured.
- The expression for each of the distance measures D_T , D_p , D_A and D_L can be expanded as a power series in z and we calculated the two leading-order terms for each of them which are determined by $H(t_o)$ and $q(t_o)$. In particular, we did this for the luminosity distance D_L as a function of z , see (107). This can be linked to observations if standard candles are available.

When using the field equation, we could establish stronger results:

- For dust solutions without a cosmological constant, we gave the relation between scale factor and time analytically in parametric form. We derived an *exact* relation between luminosity distance and redshift which is known as the Mattig relation, see (179). Just as the approximate second-order formula for an arbitrary Robertson-Walker universe, the Mattig formula is determined by $H(t_o)$ and $q(t_o)$.

- For dust solutions with a cosmological constant, we did not give the relation between scale factor and time in analytical form, although this is possible in terms of elliptic integrals. We just qualitatively discussed the influence of Λ on the scale factor. With the exact analytical solution one can derive generalised Mattig relations, see M. Dąbrowski and J. Stelmach, *Astron. J.* 92, 1272 (1986). However, we did not (and will not) work them out because they are very complicated. The distance-redshift relation in a dust universe with a cosmological constant is usually evaluated numerically. Keep in mind that it involves $\Omega_m(t_o)$ and $\Omega_\Lambda(t_o)$ and that the case $k = 0$ is characterised by the equation $\Omega_m(t_o) + \Omega_\Lambda(t_o) = 1$.

We will now link these mathematical results to observations, following the historic development.

In 1927 G. Lemaître and in 1929 E. Hubble claimed that there is a linear relation between luminosity distance and redshift. Lemaître interpreted this as evidence for an expansion of the universe. Hubble's claim was based on a sample of about 25 galaxies whose redshifts had been measured before. He used Cepheids (variable stars whose period is related to their luminosity) as standard candles in combination with some rather rough estimates. He announced that the “ K factor” (that's what we now call the Hubble constant) had a value of about 500 (km/s)/Mpc.

In the 1950s it was realised that the distance ladder, based on observations of Cepheids, had to be recalibrated. This reduced the Hubble constant $H(t_0)$, which up to this time had been generally assumed to be bigger than 200 (km/s)/Mpc, by a factor of 2. Until the 1990s, there was a controversy between two groups of cosmologists, one advocating a Hubble constant of about 50 (km/s)/Mpc and the other advocating a Hubble constant of about 100 (km/s)/Mpc. The deceleration parameter $q(t_0)$ was generally believed to be positive (corresponding to a decelerating expansion of the universe), but actually the observations were too inaccurate for determining $q(t_o)$. On theoretical grounds, many cosmologists were in favour of the Einstein-deSitter universe where the deceleration parameter is independent of time and equal to 1/2, recall Worksheet 3.

In 1998/1999 the results from two groups were published which both used supernovae of type Ia as standard candles. These supernovae are believed to be white dwarfs in a binary system. If so much mass from the companion has been accreted onto the white dwarf that its mass exceeds the Chandrasekhar limit of $1.44 M_\odot$, the white dwarf becomes unstable and explodes as a supernova. As this instability occurs at a fixed value of the mass, there is a universal relation between the shape of the light curve and the luminosity. This is why these supernovae are good standard candles. (Actually, the situation is a bit more complicated. Supernovae of type Ia are characterised by the fact that their spectra show no hydrogen lines but a silicon line. Not all supernovae of this type can be used as standard candles; one has to exclude some subclasses for which the relation between the light curve and the luminosity is different.) From observations of about 45 supernovae of type Ia in galaxies at redshifts up to $z \approx 0.9$ both groups independently found that the data *cannot* be matched to the distance-redshift relation of a universe with $q(t_0) > 0$, in particular not to a dust universe without a cosmological constant. For a dust universe with a positive cosmological constant, however, it worked, see Figure 35. If one assumes a spatially flat universe ($k = 0$), as suggested by inflation and supported by the cosmic background radiation (see below), the best fit to the supernovae Ia data suggests that the density parameters should be $\Omega_m(t_o) = 0.3$ for matter (90 % of which is assumed to be dark matter) and $\Omega_\Lambda(t_o) = 0.7$ for the cosmological constant (which may be re-interpreted as dark

energy or quintessence). The Hubble constant came out as $H(t_0) \approx 65(\text{km/s})/\text{Mpc}$; the present data, also including observations of the cosmic background radiation, are in favour of a slightly bigger value of about $H(t_0) \approx 70(\text{km/s})/\text{Mpc}$. The precise value of the deceleration parameter $q(t_0)$ is still unclear, but the supernovae Ia observations showed that it must be negative with a confidence of 7σ , see again Figure 35 where the allowed region is well above the $q(t_0) = 0$ line, i.e., in the region where $q(t_0)$ is negative. For the observation that our universe is accelerating S. Perlmutter, A. Riess and B. Schmidt won the Nobel Prize in Physics 2011. The determination of the distance-redshift relation with the help of supernovae of type Ia has been extended to redshifts bigger than 1 in the years after 2000. In addition to ground-based observations, a satellite project SNAP (SuperNova Acceleration Probe) had been proposed which later became a sub-project of WFIRST (Wide Field Infrared Survey Telescope). The future of this NASA satellite is unclear. President Trump proposed to terminate the project in February 2018 but Congress approved further funding in March 2018.

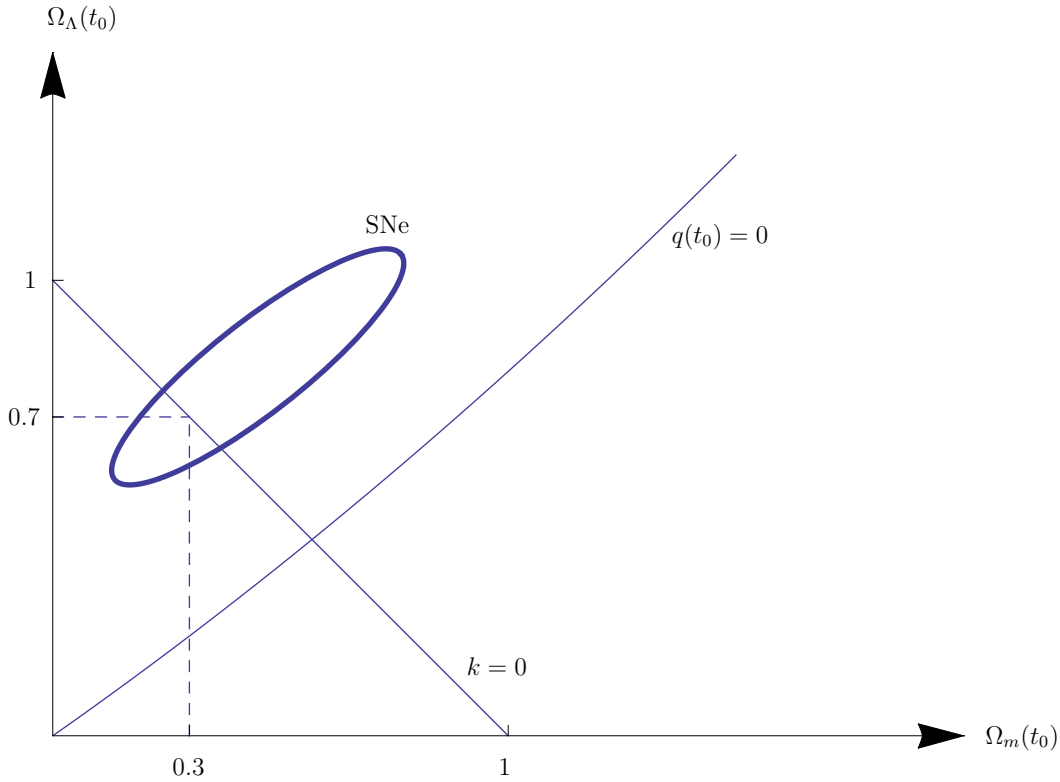


Figure 35: Restrictions on $\Omega_m(t_0)$ and $\Omega_\Lambda(t_0)$ from type Ia supernovae observations

4.3 The cosmic background radiation

We have already mentioned the chequered history of the cosmic background radiation. Recall that the officially recognised detection was made in the year 1964 by A. Penzias and R. Wilson who won the Nobel Prize in 1978. In the following years, the properties of the cosmic background radiation have been carefully investigated, in particular by several satellite missions. Note that the maximum of the cosmic background radiation is in the frequency range

of microwaves where the radiation is largely blocked by the water vapour in our atmosphere. Therefore, observations of the cosmic background radiation are made with satellites, with balloons, or with ground-based telescopes at high altitude, in particular near the South Pole where because of the cold temperature the amount of water vapour in the atmosphere is low. The most important projects have been the following:

- COBE (Cosmic Background Explorer)

This was a satellite that was launched in 1989. Data were released in 1992 and made a great impact. In particular, COBE found that the cosmic background radiation shows a perfect Planck spectrum and that it is isotropic to an extremely high degree, but it also found the first tiny deviations from isotropy. J. Mather and G. Smoot were awarded the Nobel Prize for these discoveries in 2006.

- Boomerang (Balloon Observations Of Millimetric Extragalactic Radiation and Geomagnetism)

As the name suggests, this was a balloon experiment. It flew two times, in 1998 and in 2003 and its most important result was that it detected the first acoustic peak in the power spectrum (see below) at precisely the position where it should be in a universe with $k = 0$.

- WMAP (Wilkinson Microwave Anisotropy Probe)

This was a NASA satellite that took data over the unusually long period from 2001 to 2010. It mapped the anisotropies over the whole sky, confirmed the first acoustic peak in the power spectrum and found the next ones.

- Planck

This European satellite was in operation from 2009 to 2013. Both the sensitivity and the resolution of the Planck satellite was even higher than that of WMAP, so Planck was able to determine the power spectrum to even higher values of ℓ (see below).

Another project that made big headlines was BICEP2, a telescope at the South Pole. In March 2014 the BICEP2 team announced that their investigation of the polarisation of the cosmic background radiation showed distinctive signatures from primordial gravitational waves. In the following months it was found that the measurements were correct but the interpretation was wrong. The so-called B-modes that have been observed in the polarisation had not been produced by primordial gravitational waves (at a very early stage of the universe) but rather by the influence of dust on the propagation of the photons in the cosmic background radiation (at a much later time).

We now discuss the most important observed features of the cosmic background radiation and their consequences.

(a) Planck spectrum

From elementary physics text-books we know that black-body radiation is characterised by the Planck spectrum

$$dn = \frac{8 \pi V \nu^2 d\nu}{c^3 \left(\exp\left(\frac{h\nu}{kT}\right) - 1 \right)}. \quad (257)$$

Here

$$\nu = \text{frequency, } \nu = \frac{\omega}{2\pi} = \frac{c|\vec{k}|}{2\pi} = \frac{c}{\lambda} \quad ,$$

n = photon number,

V = volume,

T = temperature,

h = Planck constant,

k = Boltzmann constant.

The corresponding spectral energy density is

$$\frac{d\varepsilon_r}{d\nu} = \frac{d\varepsilon_r}{dn} \frac{dn}{d\nu} = \frac{h\nu}{V} \frac{8\pi V \nu^2}{c^3 \left(\exp\left(\frac{h\nu}{kT}\right) - 1 \right)} = \frac{8\pi h \nu^3}{c^3 \left(\exp\left(\frac{h\nu}{kT}\right) - 1 \right)} \quad , \quad (258)$$

see the plot in Figure 36.

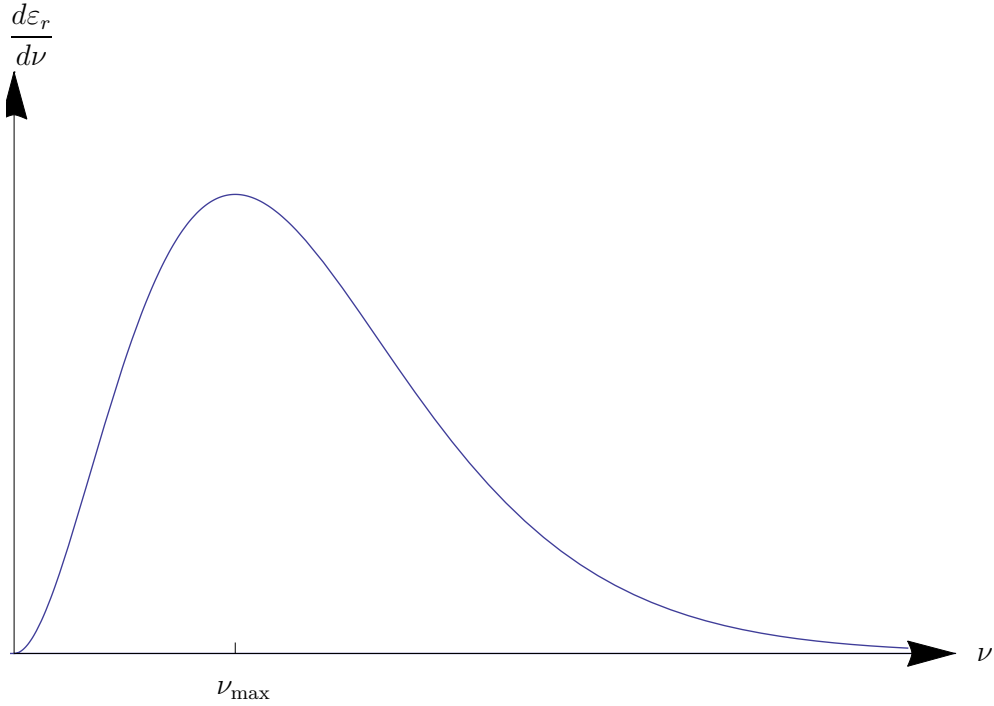


Figure 36: Spectral energy density (Planck spectrum) of black-body radiation

The maximum of the spectral energy density is at a frequency ν_{\max} which is directly proportional to T . This result is known as *Wien's displacement law*.

Integrating over all frequencies gives the *Stefan-Boltzmann law*,

$$\begin{aligned}\varepsilon_r &= \int_0^\infty \frac{d\varepsilon_r}{d\nu} d\nu = \int_0^\infty \frac{8\pi h \nu^3 d\nu}{c^3 \left(\exp\left(\frac{h\nu}{kT}\right) - 1 \right)} \\ &= \frac{8\pi k^4 T^4}{c^3 h^3} \int_0^\infty \frac{x^3 dx}{e^x - 1} = \frac{8\pi k^4 T^4}{c^3 h^3} \frac{\pi^4}{15} = \frac{4\sigma}{c} T^4\end{aligned}\quad (259)$$

where

$$\sigma = \frac{2\pi^5 k^4}{15 c^2 h^3} \approx 5.67 \times 10^{-8} \frac{\text{W}}{\text{m}^2 \text{K}^4} \quad (260)$$

is the Stefan-Boltzmann constant. This results in the equation

$$\varepsilon_r \approx 7.6 \times 10^{-16} \frac{\text{J}}{\text{m}^3} \frac{T^4}{\text{K}^4} \quad (261)$$

which allows to calculate the energy density ε_r from the temperature T .

In an expanding universe the volume at a time t_e is related to the volume at a time t_o according to

$$\frac{V(t_o)}{V(t_e)} = \frac{a(t_o)^3}{a(t_e)^3}. \quad (262)$$

With the redshift law for Robertson-Walker spacetimes,

$$\frac{\nu(t_e)}{\nu(t_o)} = 1 + z = \frac{a(t_o)}{a(t_e)}, \quad (263)$$

this can be rewritten as

$$\frac{V(t_o)}{V(t_e)} = \frac{\nu(t_e)^3}{\nu(t_o)^3}, \quad (264)$$

hence

$$V(t_o) \nu(t_o)^3 = V(t_e) \nu(t_e)^3. \quad (265)$$

Differentiation with respect to the frequency, keeping t_e and t_o fixed, yields

$$V(t_o) \nu(t_o)^2 d\nu(t_o) = V(t_e) \nu(t_e)^2 d\nu(t_e), \quad (266)$$

i.e., the numerator in the Planck law (257) is time-independent. If the photon number is conserved, this implies that also the denominator must be time-independent,

$$\frac{\nu(t_o)}{T(t_o)} = \frac{\nu(t_e)}{T(t_e)}. \quad (267)$$

Invoking again the redshift law in Robertson-Walker spacetimes, this implies

$$\frac{T(t_e)}{T(t_o)} = \frac{a(t_o)}{a(t_e)}, \quad (268)$$

i.e., if the universe expands by a certain factor, the temperature drops by the same factor, which is quite intuitive. Note that the Stefan-Boltzmann law is in agreement with the fact that

in a universe filled with radiation the energy density is inverse proportional to the fourth power of the scale factor, as we have seen before.

We observe that the cosmic background radiation reaches us now (at time t_o) with a perfect Planck spectrum whose temperature is $T(t_o) = 2.73$ K. The maximum of the radiation is at a frequency $\nu_{\max} \approx 160$ GHz which corresponds to a wavelength of $\lambda_{\max} \approx 1.1$ mm. By the Stefan-Boltzmann law, the temperature gives us the energy density $\varepsilon_r(t_o)$ of the radiation,

$$\frac{\varepsilon_r(t_o)}{c^2} \approx 4.6 \times 10^{-31} \frac{\text{kg}}{\text{m}^3}. \quad (269)$$

From the spectral distribution we can calculate that this corresponds to approximately 500 photons per cm^3 . On the other hand, the critical energy density $\varepsilon_c(t_o)$ is determined by the Hubble constant which we know rather well,

$$\frac{\varepsilon_c(t_o)}{c^2} = \mu_c(t_o) = \frac{3 H(t_o)^2}{\kappa c^4} = 1.9 \times 10^{-26} h^2 \frac{\text{kg}}{\text{m}^3} \quad (270)$$

where

$$H(t_o) = 100 \times h \frac{\text{km/s}}{\text{Mpc}}. \quad (271)$$

With $h \approx 0.7$ we find that the density parameter of the radiation is

$$\Omega_r(t_o) = \frac{\mu_r(t_o)}{\mu_c(t_o)} < 10^{-4} \quad (272)$$

which can be ignored in comparison to the density parameters of the cosmological constant and of matter, $\Omega_\Lambda(t_o) \approx 0.7$ and $\Omega_m(t_o) \approx 0.3$.

The above analysis shows that a Planck spectrum remains a Planck spectrum if photons freely propagate in an expanding universe, with the temperature being proportional to the inverse of the scale factor. The obvious idea is that the cosmic background radiation has come into existence at some time t_e , when the scale factor was smaller and the temperature was higher, and that from that time onwards the photons of the cosmic background radiation have propagated more or less freely until we observe them today at time t_o . What can we say about the time t_e ? Certainly, photons cannot propagate freely if the universe is densely filled with free electrons so that the photons undergo frequent scattering. As a rough approximation, the time t_e coincides with the time when electrons and ions formed neutral atoms, so $T(t_e)$ can be estimated from the condition that this temperature should approximately correspond to the energy where atoms are ionised,

$$k T(t_e) \approx E_{\text{ionisation}}. \quad (273)$$

A typical ionisation energy is of the order of some eV. (For hydrogen, e.g., it is 13.6 eV.) So for our rough estimate we may assume that

$$k T(t_e) \approx 1 \text{ eV}. \quad (274)$$

As the Boltzmann constant is $k = 0.86 \times 10^{-4} \text{ eV/K}$, this gives

$$T(t_e) \approx 10^4 \text{ K}. \quad (275)$$

A more detailed analysis shows that the universe became transparent at a temperature of 3000 – 4000 K. This corresponds to a redshift of $z \approx 1100$. We refer to the hypersurface $t = t_e$ with this temperature $T(t_e)$ as to the *hypersurface of last scattering*, see Figure 37. The intersection of this hypersurface with the past light-cone of an observer here and now is called the *surface of last scattering*.

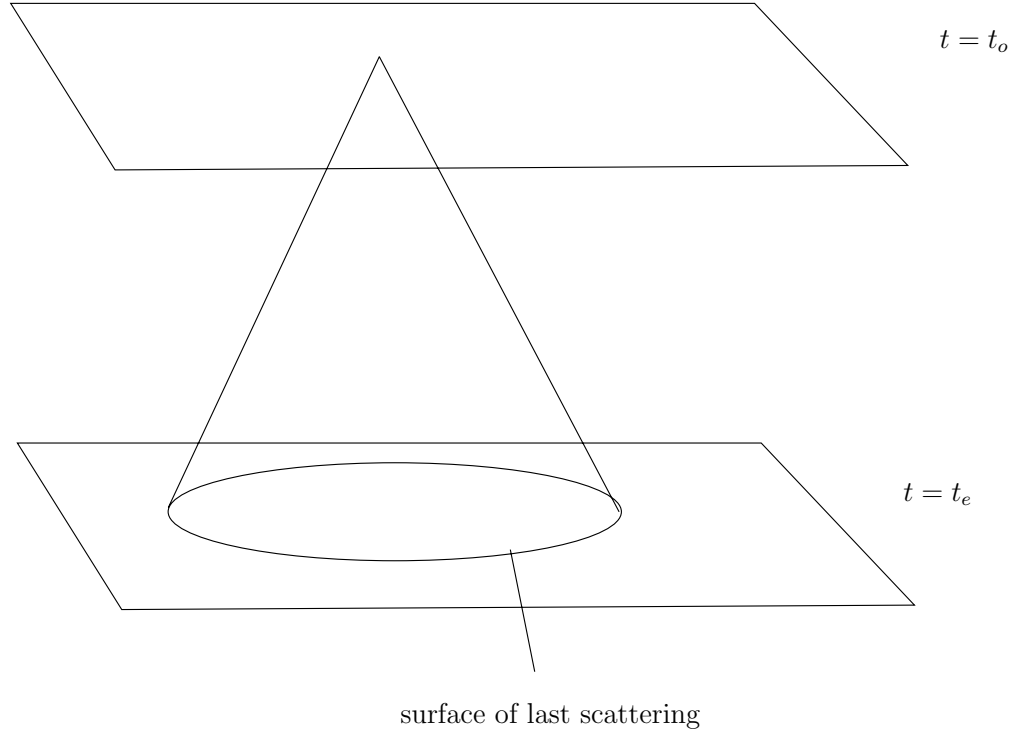


Figure 37: The hypersurface of last scattering $t = t_e$

It is usual to refer to the time when electrons and ions formed neutral atoms as to the time of *recombination*. Actually, this is a misnomer because it is for the first time in the universe that electrons and ions combine. Of course, the time of recombination was not one precise moment but rather a time interval. Similarly, the hypersurface of last scattering $t = t_e$ is an idealised model for a spacetime region with a certain temporal extension.

We have said that in the time before recombination photons underwent frequent scattering processes with free electrons. In the rest system of the electron, the photon has a certain initial energy $E_{\gamma i}$ and a certain final energy $E_{\gamma f}$. One speaks of

- Compton scattering if $E_{\gamma i} > E_{\gamma f}$,
- Thomson scattering if $E_{\gamma i} = E_{\gamma f}$,
- inverse Compton scattering if $E_{\gamma i} < E_{\gamma f}$.

In the time after recombination, the photons of the cosmic background radiation are scattered only very rarely. When they pass through the hot gas (plasma) in a galaxy cluster, these rare scattering processes can lead to a tiny distortion of the Planck spectrum. This is known as the *Sunyaev-Zel'dovich effect*.

(b) Anisotropies

The cosmic background radiation shows a Planck spectrum, so we can associate to it a temperature T . This temperature is isotropic, i.e., independent of the direction from which the radiation comes, to an extremely high degree (if we subtract the dipole term, see below). However, it is not perfectly isotropic, there are tiny anisotropies of the order of $\Delta T/T \lesssim 10^{-5}$. These tiny anisotropies give important information on the universe. They are usually modelled with the help of an expansion into spherical harmonics

$$Y_\ell^m(\vartheta, \varphi) = \sqrt{\frac{(2\ell+1)}{4\pi} \frac{(\ell-m)!}{(\ell+m)!}} P_\ell^m(\cos \vartheta) e^{im\varphi}, \quad (276)$$

where the P_ℓ^m are the associated Legendre polynomials,

$$P_\ell^m(x) = (-1)^m (1-x^2)^{m/2} \frac{d^m}{dx^m} P_\ell(x) \quad (277)$$

and the $P_\ell(x)$ are the Legendre polynomials,

$$P_\ell(x) = \frac{1}{2^\ell \ell!} \frac{d^\ell}{dx^\ell} \left((x^2-1)^\ell \right). \quad (278)$$

The temperature is a function on the sky, i.e. $T : S^2 \rightarrow \mathbb{R}^+$. The points on the sphere S^2 are in one-to-one correspondence with unit vectors \vec{e} which can be represented with the help of standard spherical polar coordinates as

$$\vec{e} = \begin{pmatrix} \sin \vartheta \cos \varphi \\ \sin \vartheta \sin \varphi \\ \cos \vartheta \end{pmatrix}. \quad (279)$$

The spherical harmonics form an orthonormal basis for square-integrable functions on the sphere, so we may write the temperature $T(\vec{e})$ as

$$T(\vec{e}) = \sum_{\ell=0}^{\infty} \sum_{m=-\ell}^{\ell} a_\ell^m Y_\ell^m(\vartheta, \varphi). \quad (280)$$

Owing to the orthonormality of the spherical harmonics with respect to the L^2 scalar product the expansion coefficients are

$$a_\ell^m = \int_0^{2\pi} \int_0^\pi T(\vartheta, \varphi) \overline{Y_\ell^m(\vartheta, \varphi)} \sin \vartheta d\vartheta d\varphi \quad (281)$$

where overlining means complex conjugation.

For each ℓ ,

$$T_\ell(\vartheta, \varphi) = \sum_{m=-\ell}^{\ell} a_\ell^m Y_\ell^m(\vartheta, \varphi) \quad (282)$$

is the multipole moment of degree ℓ of the temperature distribution.

The monopole moment ($\ell = 0$) is just the average of the temperature over the sky, so it is a constant. Observation yields

$$T_0 \approx 2.73 \text{ K}. \quad (283)$$

The dipole moment

$$T_1(\vartheta, \varphi) = \sum_{m=-1}^1 a_1^m Y_1^m(\vartheta, \varphi) \quad (284)$$

was measured around 1970. It was found that $|T_1(\vartheta, \varphi)/T_0| \lesssim 10^{-3}$. This relatively large dipole moment is understood as a result of the motion of the Earth with respect to the rest system of the cosmic background radiation: If the cosmic background radiation were perfectly isotropic with respect to the standard observers in a Robertson-Walker observer, any other observer would see a dipole anisotropy that can be explained as a Doppler effect resulting from the motion of this observer relative to the standard observers. In the forward direction the Doppler effect causes a blueshift of the photons which results in a Planck spectrum with a higher temperature, in the backward direction the Doppler effect causes a redshift of the photons which results in a Planck spectrum with a lower temperature. When we talk about anisotropies in the cosmic background radiation we always subtract the dipole term, i.e., we consider the quantity

$$\delta^T(\vec{e}) = T(\vec{e}) - T_0 - T_1(\vartheta, \varphi) = \sum_{\ell=2}^{\infty} \sum_{m=-\ell}^{\ell} a_{\ell}^m Y_{\ell}^m(\vartheta, \varphi). \quad (285)$$

Of course, one also has to subtract all other influences, in particular the emission from the galactic disc, see Figure 38. Then one finds that the temperature is isotropic to a very high degree, $|\delta^T(\vec{e})/T_0| \lesssim 10^{-5}$. It is natural to assume that δ^T is *Gaussian* (i.e., that temperature values measured in directions that are randomly distributed over the sky show a Gaussian distribution about the mean-value) and *statistically isotropic* (i.e., that the values for all higher-order multipole moments vary randomly over the sky without distinguishing a particular direction). However, the recent satellite missions WMAP and Planck, which have measured the anisotropies in the cosmic background radiation with a high accuracy, have found some indications for non-Gaussianities and also for a distinguished axis in the sky (sometimes called the “axis of evil”) with which the quadrupole moment and the octupole moment seem to be aligned. These observations have to be confirmed, so at the moment it is not yet clear if the assumptions of Gaussianity and of statistical isotropy really have to be dropped.

Here we take a conservative view and assume that Gaussianity and statistical isotropy do hold. Then the two-point autocorrelation function

$$C^T(\vartheta) = \langle \delta^T(\vec{e}) \delta^T(\vec{e}') \rangle, \quad \vec{e} \cdot \vec{e}' = \cos \vartheta, \quad (286)$$

of the temperature anisotropy δ^T determines the correlation completely because for a Gaussian distribution all higher-order correlation functions are determined by the two-point correlation function. Moreover, statistical isotropy implies that C^T may, indeed, be viewed as a function only of the angle ϑ . In (286), the pointed brackets denote an ensemble average, i.e., it is assumed that a large sample of pairs of randomly chosen directions (\vec{e}, \vec{e}') with $\vec{e} \cdot \vec{e}' = \cos \vartheta$ is taken and then the average is calculated.

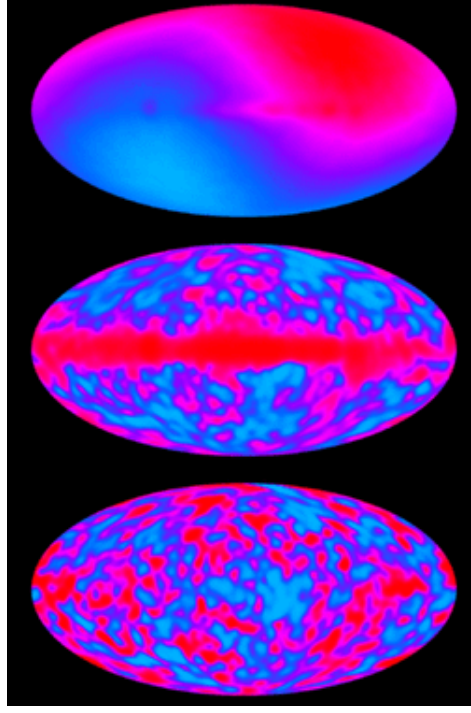


Figure 38: These pictures show the temperature anisotropies of the cosmic background radiation as measured by the COBE satellite in the 1990s. Temperatures higher than average are shown as red, temperatures lower than average are shown as blue. The first picture shows the temperature distribution just as measured; this is dominated by the dipole term. In the second picture the dipole term is subtracted. In the third picture also the emission from the galactic disc is subtracted; then a highly isotropic distribution remains. [Picture from lambda.gsfc.nasa.gov]

As C^T depends only on ϑ , expansion of this function with respect to spherical harmonics involves only terms which are independent of φ , i.e., terms with $m = 0$. As Y_ℓ^0 is a multiple of P_ℓ , this results in an expansion in terms of the Legendre polynomials,

$$C^T(\vartheta) = \sum_{\ell=0}^{\infty} \frac{2\ell+1}{4\pi} C_\ell^T P_\ell(\cos \vartheta). \quad (287)$$

The coefficients C_ℓ^T give the *angular power spectrum* of the temperature anisotropy. (This is not the general definition of the angular power spectrum, but in the case at hand it is an equivalent definition.) High values of ℓ correspond to correlations on a small angular scale. Note that because the Legendre polynomials satisfy an orthogonality condition with respect to the L^2 scalar product, the expansion equation of $C^T(\vartheta)$ can be solved for the coefficients C_ℓ^T ,

$$C_\ell^T = 2\pi \int_0^\pi C^T(\vartheta) P_\ell(\cos \vartheta) \sin \vartheta d\vartheta.$$

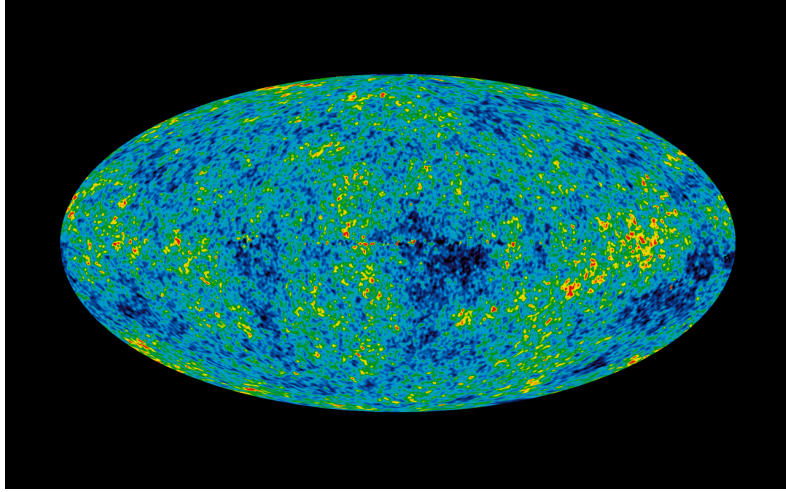


Figure 39: Temperature anisotropies in the cosmic background radiation as measured by the WMAP satellite [picture from www.nasa.gov]

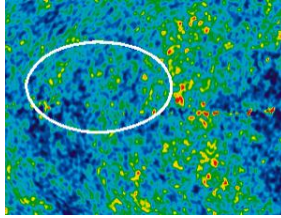


Figure 40: Initials of Stephen Hawking in the cosmic background radiation

The measurement of the C_ℓ^T for many values of ℓ was one of the main goals of the satellite missions WMAP and Planck. Although ℓ is a discrete variable, taking only non-negative integer values, C_ℓ^T is usually plotted against ℓ as if ℓ could take all non-negative real values. Figure 41 shows the results from the WMAP mission.

Already the balloon mission Boomerang had observed a local maximum of C_ℓ^T near $\ell = 200$. WMAP and Planck confirmed this and found additional local maxima at higher values of ℓ . Figure 41 clearly shows at least the second one. These local maxima are known as “acoustic peaks”. Before they were observed, they had actually been predicted on the basis of the following theoretical consideration: In the era before recombination, ions, electrons and photons formed a kind of soup where with a certain statistical probability overdensities formed. Each overdensity grew over a certain time, because it gravitationally attracted neighbouring matter, until the pressure became so big that a (roughly spherical) wave expanded from the overdensity. This is quite similar to the formation of a sound wave in a gas, therefore one calls these waves “acoustic”. The distance the photons, which are part of the soup, could travel before they decoupled from the matter at about the time of recombination, depends on the speed of sound. The latter can be theoretically calculated with the help of perturbation theory, see the next chapter. After the time of recombination, the photons decouple from the matter and just freely follow the expansion of the universe. This process results in the formation of roughly spherical shells of photons with a radius that can be theoretically predicted. Clearly, the existence of such shells results in a certain correlation of the anisotropy of the cosmic background radiation at a certain angular scale, with a maximum at a particular value of ℓ . On the basis of a

universe with $k = 0$, as suggested by inflation, the first acoustic peak was predicted to occur near $\ell = 200$. This was precisely what the observations have shown. Also the discovery of the other peaks is in agreement with the assumption that $k = 0$. As these calculations are quite sensitive to the value of k , the location of the acoustic peaks give strong support to the idea that we live in a universe with $k = 0$.

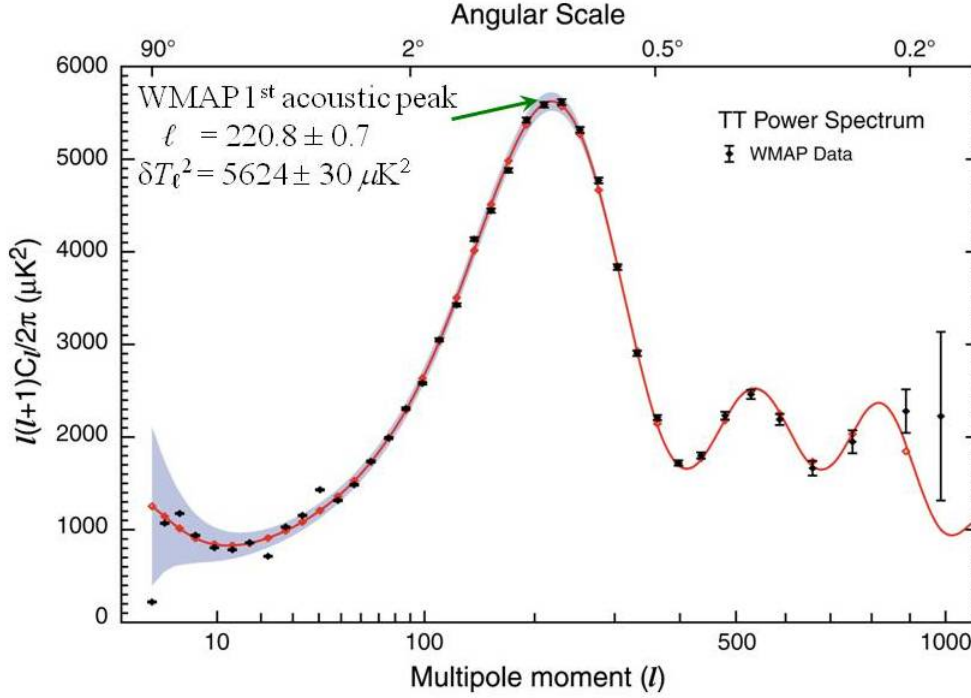


Figure 41: Angular power spectrum of the cosmic background radiation as measured by WMAP [picture from telescoper.wordpress.com]

Recall that the observations of the supernovae of type Ia could be explained by assuming a universe with a cosmological constant and a dust, i.e., with two density parameters $\Omega_\Lambda(t_o)$ and $\Omega_m(t_o)$. The data located the values for these density parameters within an elliptical area, see Figure 35. If we combine this result with the evidence for $k = 0$, as it comes from the anisotropy of the cosmic background radiation, we have to intersect this elliptical area with the straight line where $\Omega_\Lambda(t_o) + \Omega_m(t_o) = 1$ which corresponds to $k = 0$. This gives the values of approximately $\Omega_\Lambda(t_o) = 0.7$ and $\Omega_m(t_o) = 0.3$.

So our present understanding that we live in a universe with $k = 0$, $\Omega_\Lambda(t_o) = 0.7$, $\Omega_m(t_o) = 0.3$ and $\Omega_r(t_o)$ negligibly small is based on a combination of observations of the distance-redshift relation and of the cosmic background radiation. There is still a certain controversy about the precise value of the Hubble constant $H(t_o)$. As mentioned above, the measurement of the distance-redshift relation with the help of type Ia supernovae gave a value of $H(t_o) \approx 65 \text{ (km/s)/Mpc}$. The data from WMAP and Planck indicated that the value of the Hubble constant is actually bigger, $H(t_o) \approx 70 \text{ (km/s)/Mpc}$.

(c) Polarisation

The cosmic background radiation is unpolarised to a very high degree. Relative deviations are of the order $\lesssim 10^{-6}$ which is even one order of magnitude smaller than that for the temperature anisotropy. Nonetheless, a slight degree of polarisation has been detected.

For a theoretical description, one uses again an expansion into spherical harmonics. However, this is now much more involved than for the temperature because the degree of polarisation cannot be described by a scalar variable: One usually uses the so-called Stokes parameters which can be combined to form a second-rank tensor field on the celestial sphere. Therefore, one cannot expand the polarisation measure in terms of the usual *scalar* spherical harmonics, one rather has to use *tensorial* spherical harmonics. There are two families of such tensorial spherical harmonics, one of them for the expansion of curl-free anisotropies and one for divergence-free anisotropies. Because of the analogy with electrodynamics, the curl-free anisotropies are called E-modes and the divergence-free anisotropies are called B-modes. (Note that this has nothing to do with the real electric or magnetic fields of which the cosmic background radiation consists!) Theoretically the slight degree of polarisation of the cosmic background radiation can be explained by the influence of matter on the photons on their way from the surface of last scattering to us, i.e., as scattering or deflection (lensing) effects. E-modes that could be explained in this way have been observed for the first time in 2003 and B-modes that could be explained in this way have been observed in 2013. In 2014, it was announced that the telescope BICEP2 at the South Pole had observed a kind of B-modes that could be explained only as the result of primordial gravitational waves; this would have been a kind of anisotropy imprinted on the cosmic background radiation already when it left the surface of last scattering. If true, this would have given strong support for the idea of inflation because otherwise it would have been impossible to explain how the effect of primordial gravitational waves could have grown to a measurable size. However, it was found out that the BICEP2 observations could very well be explained as the effect of dust (“foreground”) on the cosmic background radiation. The BICEP2 team withdrew the announcement of having detected primordial gravitational waves after a few months. Note that it is generally accepted that the *observations* of the BICEP2 team were correct; it is the *interpretation* of these observations that was wrong.

4.4 Other observations

Without going into details, we very briefly indicate that our cosmological models are also restricted by some other kind of observations.

- Number counts

Let us assume we count all galaxies in the sky up to a certain magnitude, i.e., all galaxies whose flux is bigger than a certain chosen limit value F . How does the number N of these galaxies depend on the flux F ? In a Euclidean static universe, N would grow with R^3 where R is the radius of the volume in which we count the galaxies. On the other hand, the flux falls off with R^{-2} , so we have $N \sim F^{-3/2}$. As the log of F gives the magnitude, one usually writes this as $\log N = -(3/2) \log F + \text{constant}$. For an expanding (and possibly non-Euclidean) universe, we get a different relation. In this way number counts give us a means for testing a chosen cosmological model.

Unfortunately, this method is not very reliable. The reason is that galaxies develop over time. On average, a distant galaxy is seen at a younger stage of its life than a galaxy closer by. We do not know enough about the development of galaxies for accurately estimating the effect of age on the intrinsic luminosity.

- Baryonic Acoustic Oscillations (BAO)

We have briefly mentioned the formation of acoustic peaks in the cosmic background radiation from spherical acoustic waves that formed at a time before recombination. Not only the photons take part in these acoustic waves, but also the baryons. So they should also form a spherical shell about each centre where an overdensity had formed. This should be visible in the two-point correlation function *for the matter density*. The Sloan Digital Sky Survey has revealed indications for these so-called Baryonic Acoustic Oscillations. They are in agreement with the Λ CDM model with $\Omega_\Lambda(t_o) = 0.7$ and $\Omega_m(t_o) = 0.3$.

- Gravitational lensing

Microlensing plays an important role for estimating the dark matter that can exist in the form of Massive Compact Halo Objects and weak lensing is crucial for estimating the dark matter in galaxy clusters. This was outlined already in Section 4.1. In addition, lensing is also relevant for determining the matter in the universe at very large scales. The same kind of weak lensing observations that has been made in the direction of galaxy clusters has also been made in directions where no galaxy clusters are visible. Any deviation from a random distribution of the shapes of background galaxies would indicate a deforming influence of the matter distribution in the universe on the cross-sections of light bundles at very large scales. This so-called “cosmic shear” was detected around the year 2000. It restricts the possible ways in which we can model our universe as a Robertson-Walker universe with certain perturbations. As the weak-lensing observations can only determine the surface mass density (i.e., mass per area of a surface perpendicular to the line of sight) it cannot determine a 3D distribution of matter in the universe. However, if weak lensing is combined with other observations, it is possible to produce 3D maps of the distribution of matter. These maps show a strong tendency of the matter to form filaments at very large scales, see Figure 42.

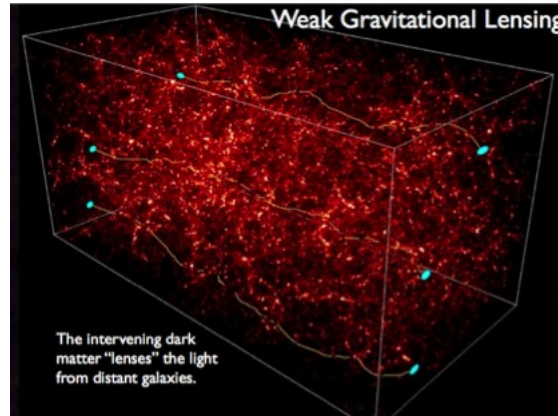


Figure 42: Large-scale 3D distribution of matter as constructed from weak-lensing observations [picture by S. Colombi and Y. Mellier, taken here from lateuniverse.wordpress.com]

5 Perturbation theory

If we want to theoretically describe the anisotropies in the cosmic background radiation, and other anisotropies, we have to go beyond the homogeneous and isotropic cosmological models, i.e., beyond Robertson-Walker spacetimes. Cosmological perturbation theory is the usual mathematical setting for this kind of investigations.

General relativistic perturbation theory is based on an ansatz for the metric of the form

$$g_{\mu\nu} = \bar{g}_{\mu\nu} + \delta g_{\mu\nu} \quad (288)$$

where $\bar{g}_{\mu\nu}$ is a given (“background”) metric. One assumes that the perturbation is so small that it is justified to linearise all equations with respect to $\delta g_{\mu\nu}$ and its derivatives. In this way, the nonlinear Einstein equation for the metric $g_{\mu\nu}$ is reduced to a linear differential equation for the perturbation $\delta g_{\mu\nu}$. This linearised formalism is best known for the case that $\bar{g}_{\mu\nu}$ is the Minkowski metric. In an appropriate gauge (i.e., if the freedom of making coordinate transformations is used in an intelligent way), the resulting vacuum equation for $\delta g_{\mu\nu}$ reduces to the ordinary wave equation and thus to the Laplace equation for the static case. In this setting Einstein derived the perihelion precession of Mercury, the light deflection at the Sun and the existence of gravitational waves. The linearised formalism is also well developed for the case that $\bar{g}_{\mu\nu}$ is the Schwarzschild metric. After decomposing the perturbation into two parts that transform differently under spatial inversion (parity), this leads to the Regge-Wheeler equation for one part and to the Zerilli equation for the other.

In cosmological perturbation theory, it is natural to choose $\bar{g}_{\mu\nu}$ to be a Robertson-Walker metric. This formalism dates back to a pioneering paper by Y. Lifshitz (1946), but it developed into a powerful tool only after J. Bardeen (1980) wrote the perturbation functions in a way that is invariant under coordinate transformations. As in perturbation theory a change of coordinates is somewhat similar to a gauge transformation in electrodynamics, it is usual to refer to Bardeen’s formalism as to *gauge-invariant* perturbation theory. We will now explain the basic features of this formalism.

For the sake of simplicity, we restrict to the case that the background metric is a *spatially flat* Robertson-Walker universe, i.e., to the case $k = 0$, with spatial topology \mathbb{R}^3 . Then

$$\bar{g}_{\mu\nu} dx^\mu dx^\nu = -c^2 dt^2 + a^2 \left(d\chi^2 + \chi^2 (d\vartheta^2 + \sin^2 \vartheta d\varphi^2) \right) \quad (289)$$

where the scale factor a is a function of t . If we use the conformal time η , this metric can be rewritten as

$$\bar{g}_{\mu\nu} dx^\mu dx^\nu = a^2 \left(-d\eta^2 + d\chi^2 + \chi^2 (d\vartheta^2 + \sin^2 \vartheta d\varphi^2) \right) \quad (290)$$

where now a has to be viewed as a function of η . What we have in the brackets is just the Minkowski metric in spherical polars, so we may rewrite it in terms of Cartesian coordinates as

$$\bar{g}_{\mu\nu} dx^\mu dx^\nu = a^2 \left(-d\eta^2 + \delta_{ij} dx^i dx^j \right). \quad (291)$$

As usual, latin indices i, j, \dots take values 1,2,3 and, for this section, we agree to lower and to raise latin indices with δ_{ij} and δ^{ij} , respectively. The metric is now, up to the conformal factor a^2 , the Minkowski metric in usual inertial coordinates. Note, however, that η and χ

are dimensionless while usually, when writing the Minkowski metric in spherical polars, we use $x^0 = ct$ and r which have the dimension of a length. So we have to keep in mind that the “Minkowski-like” coordinates η , x^1 , x^2 and x^3 are dimensionless and that the correct dimension of $\bar{g}_{\mu\nu}$ is provided by the conformal factor a^2 which has the dimension of a length.

We now switch on the perturbation. If we label the components of $\delta g_{\mu\nu}$ appropriately, this can be written as

$$g_{\eta\eta} = \bar{g}_{\eta\eta} + \delta g_{\eta\eta} = -a^2 + \delta g_{\eta\eta} = -a^2 (1 + 2A), \quad (292)$$

$$g_{\eta i} = \bar{g}_{\eta i} + \delta g_{\eta i} = 0 + \delta g_{\eta i} = a^2 B_i, \quad (293)$$

$$g_{ij} = \bar{g}_{ij} + \delta g_{ij} = a^2 \delta_{ij} + \delta g_{ij} = a^2 (\delta_{ij} + h_{ij}). \quad (294)$$

Then the perturbed metric reads

$$g_{\mu\nu} dx^\mu dx^\nu = a^2 \left(- (1 + 2A) d\eta^2 + 2 B_i dx^i d\eta + (\delta_{ij} + h_{ij}) dx^i dx^j \right). \quad (295)$$

Here it is convenient to further decompose the vectorial part, B_i , and the tensorial part, h_{ij} , of the perturbation. As we are on a spatially flat background, we can use ordinary vector calculus. It is well known that any vector field on Euclidean 3-space can be decomposed into a curl-free and a divergence-free vector field. This is known as the *Helmholtz decomposition theorem*, so we may write in the case at hand

$$B_i = \partial_i B + \hat{B}_i \quad (296)$$

where B is a scalar field and \hat{B}_i is divergence-free, $\partial_i \hat{B}^i = 0$.

Sketch of proof of the Helmholtz decomposition theorem:

We want to write a given vector field \vec{B} in the form

$$\vec{B} = \vec{\nabla} B + \vec{\tilde{B}} \quad (297)$$

where $\vec{\nabla} \cdot \vec{\tilde{B}} = 0$. If this equation holds, we must have

$$\vec{\nabla} \cdot \vec{B} = \Delta B. \quad (298)$$

For any smooth \vec{B} , this equation has a solution which is, of course, not unique. A particular solution may be written as

$$B(\vec{x}) = -\frac{1}{4\pi} \int_{\mathbb{R}^3} \frac{(\vec{\nabla}' \cdot \vec{B})(\vec{x}')}{|\vec{x} - \vec{x}'|} d^3 \vec{x}', \quad (299)$$

as is well known from electrodynamics, provided that \vec{B} falls off sufficiently strongly so that the integral exists. In any case, with a chosen solution B we may *define* $\vec{\tilde{B}} := \vec{B} - \vec{\nabla} B$. Then $\vec{\nabla} \cdot \vec{\tilde{B}} = 0$ and we are done.

Similarly to the Helmholtz decomposition of vector fields, one may also decompose tensor fields of second rank. We just give the result here, for details we refer to J. Ehlers' notes in General Relativity and Gravitation 39, 1929 (2007). One finds that the symmetric second-rank tensor field h_{ij} can be written as

$$h_{ij} = 2C\delta_{ij} + \left(\partial_i\partial_j - \frac{1}{3}\delta_{ij}\Delta\right)E + \partial_i\hat{E}_j + \partial_j\hat{E}_i + 2\hat{E}_{ij} \quad (300)$$

where C and E are scalar fields, \hat{E}_i is a vector field with $\partial_i\hat{E}^i = 0$ and \hat{E}_{ij} is a second-rank tensor field with $\partial_i\hat{E}^{ij} = 0$ and $\hat{E}_i{}^i = 0$. Following a general convention, we denote tensor fields that are divergence-free by a hat. This puts the perturbed metric into the following form:

$$g_{\mu\nu}dx^\mu dx^\nu = a^2 \left\{ -\left(1 + 2A\right)d\eta^2 + \left(\partial_i B + \hat{B}_i\right)dx^i d\eta + \left((1 + 2C)\delta_{ij} + \left(\partial_i\partial_j - \frac{1}{3}\delta_{ij}\Delta\right)E + 2\partial_i\hat{E}_j + 2\hat{E}_{ij}\right)dx^i dx^j \right\}. \quad (301)$$

We have just relabelled the perturbation: In the beginning we had the $\delta g_{\mu\nu}$ which form a symmetric 4×4 matrix, i.e., there are 10 independent scalar perturbation functions. After the relabelling we have

- 4 scalar fields A, B, C, E ,
- two (co)vector fields \hat{B}_i and \hat{E}_i ,
- one symmetric second-rank tensor field \hat{E}_{ij}

which have $4 + 6 + 6 = 16$ scalar components.

They are restricted by the constraints

- $\partial_i\hat{E}^i = 0$, $\partial_i\hat{B}^i = 0$
- $\partial_i\hat{E}^{ij} = 0$,
- $\hat{E}_i{}^i = 0$

which are $2 + 3 + 1 = 6$ conditions. So altogether we have $16 - 6 = 10$ independent scalar perturbation variables which is indeed the same number as before.

The perturbations $\partial_i B$ and $\partial_i E$ are often called *longitudinal*, while the perturbations \hat{B}_i , \hat{E}_i and \hat{E}_{ij} are called *transverse*. This terminology refers to Fourier transformations: If we expand all terms with respect to the spatial variables as integrals $\int \dots \exp(k_i x^i) d^3\vec{k}$, there are terms proportional to \vec{k} and terms perpendicular to \vec{k} . For obvious reasons, the former are called “longitudinal” while the latter are called “transverse”. Note that Fourier expansion requires square integrability of the perturbations which is usually assumed to hold in cosmological perturbation theory.

You may ask what is the advantage of relabelling the perturbation in such a rather complicated form. The answer is that in terms of the new variables $A, B, C, E, \hat{E}_i, \hat{B}_i, \hat{E}_{ij}$ it is easier to find out which perturbations are gauge invariant and to decompose a general perturbation into scalar, vector and second-rank tensor parts that have a coordinate-independent meaning.

To make this clear, we have to investigate how the perturbations transform under coordinate changes. As we are interested in perturbations only to within the linear approximation, it suffices to consider coordinate transformations $(\eta, x^1, x^2, x^3) \mapsto (\tilde{\eta}, \tilde{x}^1, \tilde{x}^2, \tilde{x}^3)$ of the form

$$\tilde{\eta} = \eta + \tau, \quad \tilde{x}^i = x^i + \xi^i = x^i + \partial^i \xi + \hat{\xi}^i \quad (302)$$

where τ and ξ^i are assumed to be so small that all equations can be linearised with respect to them and their derivatives. In (302) we have decomposed ξ^i according to the Helmholtz theorem into the gradient of a scalar field ξ and a vector field $\hat{\xi}^i$ with $\partial_i \hat{\xi}^i = 0$.

We have now to calculate how the metric coefficients transform under such a coordinate transformation. We work this out in detail for the time-time component:

$$\begin{aligned} g\left(\frac{\partial}{\partial\eta}, \frac{\partial}{\partial\eta}\right) &= g\left(\frac{\partial\tilde{\eta}}{\partial\eta} \frac{\partial}{\partial\tilde{\eta}} + \frac{\partial\tilde{x}^i}{\partial\eta} \frac{\partial}{\partial\tilde{x}^i}, \frac{\partial\tilde{\eta}}{\partial\eta} \frac{\partial}{\partial\tilde{\eta}} + \frac{\partial\tilde{x}^j}{\partial\eta} \frac{\partial}{\partial\tilde{x}^j}\right) \\ &= \left(\frac{\partial\tilde{\eta}}{\partial\eta}\right)^2 g\left(\frac{\partial}{\partial\tilde{\eta}}, \frac{\partial}{\partial\tilde{\eta}}\right) + 2 \frac{\partial\tilde{\eta}}{\partial\eta} \frac{\partial\tilde{x}^i}{\partial\eta} g\left(\frac{\partial}{\partial\tilde{\eta}}, \frac{\partial}{\partial\tilde{x}^i}\right) + \frac{\partial\tilde{x}^i}{\partial\eta} \frac{\partial\tilde{x}^j}{\partial\eta} g\left(\frac{\partial}{\partial\tilde{x}^i}, \frac{\partial}{\partial\tilde{x}^j}\right), \end{aligned} \quad (303)$$

$$g_{\eta\eta} = \left(1 + \frac{\partial\tau}{\partial\eta}\right)^2 \tilde{g}_{\eta\eta} + 2 \left(1 + \frac{\partial\tau}{\partial\eta}\right) \frac{\partial\xi^i}{\partial\eta} \tilde{g}_{\eta i} + \frac{\partial\xi^i}{\partial\eta} \frac{\partial\xi^j}{\partial\eta} \tilde{g}_{ij}. \quad (304)$$

The terms $\partial\xi^i/\partial\eta$ and $\tilde{g}_{\eta i}$ are of first order in the perturbations, so the second and the third term on the right-hand side of (304) are of second order and thus negligible. Hence

$$\begin{aligned} g_{\eta\eta} &= \left(1 + \frac{\partial\tau}{\partial\eta}\right)^2 \tilde{g}_{\eta\eta}, \\ -a(\eta)^2 (1 + 2A) &= -\left(1 + \frac{\partial\tau}{\partial\eta}\right)^2 a(\eta + \tau)^2 (1 + 2\tilde{A}), \\ a(\eta)^2 (1 + 2A) &= \left(1 + 2\frac{\partial\tau}{\partial\eta} + \dots\right) \left(a(\eta) + \frac{da(\eta)}{d\eta} \tau + \dots\right)^2 (1 + 2\tilde{A}) \\ &= \left(1 + 2\frac{\partial\tau}{\partial\eta} + \dots\right) a(\eta)^2 \left(1 + \frac{2}{a(\eta)} \frac{da(\eta)}{d\eta} \tau + \dots\right) (1 + 2\tilde{A}) \\ &= a(\eta)^2 \left(1 + 2\frac{\partial\tau}{\partial\eta} + \frac{2}{a(\eta)} \frac{da(\eta)}{d\eta} \tau + 2\tilde{A} + \dots\right), \\ 1 + 2A &= 1 + 2\frac{\partial\tau}{\partial\eta} + \frac{2}{a(\eta)} \frac{da(\eta)}{d\eta} \tau + 2\tilde{A}, \end{aligned}$$

and thus

$$A = \frac{\partial\tau}{\partial\eta} + \mathcal{H}\tau + \tilde{A}, \quad (305)$$

where \mathcal{H} is the Hubble “constant” with respect to conformal time,

$$\mathcal{H}(\eta) = \frac{1}{a(\eta)} \frac{da(\eta)}{d\eta}. \quad (306)$$

Similarly, the transformation of all the other metric coefficients can be calculated. If we solve for the twiddled quantities, we find

$$\tilde{A} = A - \frac{\partial \tau}{\partial \eta} - \mathcal{H} \tau, \quad (307)$$

$$\tilde{B} = B + \tau - \frac{\partial \xi}{\partial \eta}, \quad (308)$$

$$\tilde{C} = C - \mathcal{H} \tau - \frac{1}{3} \Delta \xi, \quad (309)$$

$$\tilde{E} = E - \xi, \quad (310)$$

$$\tilde{\hat{B}}_i = \hat{B}_i - \frac{d\hat{\xi}_i}{d\eta}, \quad (311)$$

$$\tilde{\hat{E}}_i = \hat{E}_i - \xi_i, \quad (312)$$

$$\tilde{\hat{E}}_{ij} = \hat{E}_{ij}. \quad (313)$$

We see that by choosing the coordinate transformation (i.e., the scalar functions τ and ξ and the vector field $\hat{\xi}^i$) appropriately, we can achieve that

$$B = 0, \quad E = 0, \quad \hat{E}_i = 0, \quad (314)$$

or, alternatively,

$$B = 0, \quad E = 0, \quad \hat{B}_i = 0. \quad (315)$$

The latter choice has the advantage that the hypersurfaces $\eta = \text{constant}$ are then perpendicular to the η -lines, even in the perturbed spacetime. This is known as the *synchronous gauge*. Mixed spatial-temporal components (i.e., components $g_{\eta i}$ in our setting) are usually called *gravitomagnetic terms*. This refers to an analogy to electromagnetism: Whereas rotating charges produce magnetic fields, rotating masses produce $g_{\eta i}$ terms. Our observation that B and \hat{B}_i can be transformed to zero by a coordinate transformation implies that, within cosmological perturbation theory, gravitomagnetic terms can be gauged away.

Out of the perturbation variables $A, B, C, E, \hat{E}_i, \hat{B}_i$ and \hat{E}_{ij} we can form the following *gauge-invariant* variables which were introduced by J. Bardeen in 1980:

$$\Psi = A - \mathcal{H} \left(B - \frac{\partial E}{\partial \eta} \right) + \frac{\partial}{\partial \eta} \left(B - \frac{\partial E}{\partial \eta} \right), \quad (316)$$

$$\Phi = -C - \mathcal{H} \left(B - \frac{\partial E}{\partial \eta} \right) + \frac{1}{3} \Delta E, \quad (317)$$

$$\hat{\Phi}_i = \frac{\partial \hat{E}_i}{\partial \eta} - \hat{B}_i, \quad (318)$$

$$\hat{E}_{ij}. \quad (319)$$

It is straight-forward to verify that these quantities are indeed unchanged, to within linear approximation, under a coordinate transformation. We may work in coordinates where $B = 0$, $E = 0$ and $\hat{E}_i = 0$ and express the metric perturbations in terms of the Bardeen variables $\Psi = A$, $\Phi = -C$, $\hat{\Phi}_i = -\hat{B}_i$ and \hat{E}_{ij} , i.e.

$$g_{\mu\nu} dx^\mu dx^\nu = a^2 \left(- (1 + 2\Psi) d\eta^2 - 2 \hat{\Phi}_i dx^i d\eta + ((1 - 2\Phi)\delta_{ij} + 2 \hat{E}_{ij}) dx^i dx^j \right). \quad (320)$$

In this way we work in a particular coordinate system, but the perturbation variables have a gauge-invariant meaning. Keep in mind that a depends only on η whereas the perturbation variables depend on all four coordinates η , x^1 , x^2 and x^3 .

A fairly large part of cosmological perturbation theory restricts to scalar perturbations, i.e., to the case that only the two scalar Bardeen potentials Ψ and Φ are non-zero. In this restricted formalism we cannot, of course, describe gravitational waves, because this would require tensor perturbations $\hat{E}_{ij} \neq 0$, but we may describe e.g. density inhomogeneities.

We will now work out the linearised field equation for the case of scalar perturbations,

$$g_{\mu\nu} dx^\mu dx^\nu = a^2 \left(- (1 + 2\Psi) d\eta^2 + (1 - 2\Phi)\delta_{ij} dx^i dx^j \right). \quad (321)$$

We are interested in perfect-fluid solutions, so we have to find out on what conditions the field equation holds with an energy-momentum tensor of the form

$$T_{\rho\sigma} = \left(\varepsilon + p \right) \frac{U_\rho U_\sigma}{c^2} + p g_{\rho\sigma}. \quad (322)$$

According to the rules of linear perturbation theory, we assume that

$$U_\rho = \bar{U}_\rho + \delta U_\rho = N \delta_\rho^\eta + \delta U_\rho, \quad (323)$$

$$\varepsilon = \bar{\varepsilon} + \delta\varepsilon, \quad (324)$$

$$p = \bar{p} + \delta p, \quad (325)$$

where the overlined quantities refer to the perfect fluid associated with the unperturbed Robertson-Walker spacetime. The normalisation condition of the four-velocity requires

$$\begin{aligned} -c^2 &= g^{\rho\sigma} U_\rho U_\sigma = g^{\eta\eta} (U_\eta)^2 + g^{ij} U_i U_j = \frac{1}{-a^2(1+2\Psi)} (N^2 + 2N \delta U_\eta) + \dots \\ &= \frac{(1-2\Psi+\dots)}{-a^2} (N^2 + 2N \delta U_\eta) + \dots = -\frac{N^2}{a^2} + \frac{2\Psi N^2}{a^2} - \frac{2N \delta U_\eta}{a^2}. \end{aligned} \quad (326)$$

Comparing zeroth order terms and first order terms yields

$$N = -c a, \quad \delta U_\eta = -c a \Psi, \quad (327)$$

i.e.,

$$U_\rho = -\delta_\rho^\eta a c (1 + \Psi) + \delta U_i \delta_\rho^i. \quad (328)$$

We have chosen N negative because we want to have the vector $U^\sigma = g^{\sigma\rho} U_\rho$ future-pointing.

The components of the linearised energy-momentum tensor are

$$\begin{aligned} T_{\eta\eta} &= (\varepsilon + p) \frac{U_\eta^2}{c^2} + p g_{\eta\eta} \\ &= (\bar{\varepsilon} + \bar{p} + \delta\varepsilon + \delta p) \frac{1}{c^2} a^2 c^2 (1 + 2\Psi) - (\bar{p} + \delta p) a^2 (1 + 2\Psi) \\ &= a^2 (\bar{\varepsilon} + \delta\varepsilon + 2\bar{\varepsilon}\Psi), \end{aligned} \quad (329)$$

$$T_{\eta i} = (\varepsilon + p) \frac{U_\eta U_i}{c^2} + p g_{\eta i} = (\bar{\varepsilon} + \bar{p}) \frac{a}{c} \delta U_i, \quad (330)$$

$$T_{ij} = (\varepsilon + p) \frac{U_i U_j}{c^2} + p g_{ij} = (\bar{p} + \delta p - 2\bar{p}\Phi) a^2 \delta_{ij}. \quad (331)$$

We now have to calculate the left-hand side of the field equation

$$G_{\mu\nu} + \Lambda g_{\mu\nu} = \kappa T_{\mu\nu}. \quad (332)$$

With the help of Mathematica (or some other computer programme) we calculate for the metric (321) the Einstein tensor $G_{\mu\nu} = R_{\mu\nu} - R g_{\mu\nu}/2$ to within linear order with respect to Ψ , Φ and their derivatives. We find

$$G_{\eta\eta} = 3\mathcal{H}^2 + 2\Delta\Phi - 6\mathcal{H} \frac{\partial\Phi}{\partial\eta}, \quad (333)$$

$$G_{\eta i} = 2\partial_i \left(\frac{\partial\Phi}{\partial\eta} + \mathcal{H}\Psi \right), \quad (334)$$

$$\begin{aligned} G_{ij} &= \left(-2 \frac{d\mathcal{H}}{d\eta} - \mathcal{H}^2 + \Delta(\Psi - \Phi) + 2 \frac{\partial^2\Phi}{\partial\eta^2} + 2 \left(2 \frac{d\mathcal{H}}{d\eta} + \mathcal{H}^2 \right) (\Phi + \Psi) \right. \\ &\quad \left. + 2\mathcal{H} \frac{\partial\Psi}{\partial\eta} + 4\mathcal{H} \frac{\partial\Phi}{\partial\eta} \right) \delta_{ij} + \partial_i \partial_j (\Phi - \Psi). \end{aligned} \quad (335)$$

Comparing first-order terms on both sides of Einstein's field equation results, for the $\eta\eta$, ηi and ij components in

$$2\Delta\Phi - 6\mathcal{H} \frac{\partial\Phi}{\partial\eta} - 2\Lambda a^2 \Psi = \kappa a^2 (\delta\varepsilon + 2\bar{\varepsilon}\Psi), \quad (336)$$

$$2\partial_i \left(\frac{\partial\Phi}{\partial\eta} + \mathcal{H}\Psi \right) = \kappa (\bar{\varepsilon} + \bar{p}) \frac{a}{c} \delta U_i, \quad (337)$$

$$\begin{aligned} & \left(\Delta(\Psi - \Phi) + 2\frac{\partial^2 \Phi}{\partial \eta^2} + 2\left(2\frac{\partial \mathcal{H}}{\partial \eta} + \mathcal{H}^2\right)(\Phi + \Psi) + 2\mathcal{H}\frac{\partial \Psi}{\partial \eta} + 4\mathcal{H}\frac{\partial \Phi}{\partial \eta} \right) \delta_{ij} \\ & + \partial_i \partial_j (\Phi - \Psi) - 2\Lambda a^2 \Phi \delta_{ij} = \kappa a^2 (\delta p - 2\bar{p}\Phi) \delta_{ij}, \end{aligned} \quad (338)$$

respectively. Eq. (338) requires

$$\partial_i \partial_j (\Phi - \Psi) = 0 \quad \text{for } i \neq j. \quad (339)$$

Solving this differential equation for all index combinations $i, j = 1, 2, 3$ demonstrates that $\Phi - \Psi$ must be of the form

$$(\Phi - \Psi)(\eta, x^1, x^2, x^3) = f_1(\eta, x^1) + f_2(\eta, x^2) + f_3(\eta, x^3). \quad (340)$$

If we require the perturbations to be square-integrable over \mathbb{R}^3 , to allow for spatial Fourier expansion, this implies that

$$\Phi - \Psi = 0, \quad (341)$$

so we have only one Bardeen potential $\Phi = \Psi$ for perfect-fluid solutions. (336), (337) and (338) reduce to

$$2\Delta\Phi - 6\mathcal{H}\frac{\partial\Phi}{\partial\eta} - 2\Lambda a^2\Phi = \kappa a^2(\delta\varepsilon + 2\bar{\varepsilon}\Phi), \quad (342)$$

$$2\partial_i\left(\frac{\partial\Phi}{\partial\eta} + \mathcal{H}\Phi\right) = \kappa(\bar{\varepsilon} + \bar{p})\frac{a}{c}\delta U_i, \quad (343)$$

$$2\frac{\partial^2\Phi}{\partial\eta^2} + 4\left(2\frac{\partial\mathcal{H}}{\partial\eta} + \mathcal{H}^2\right)\Phi + 6\mathcal{H}\frac{\partial\Phi}{\partial\eta} - 2\Lambda a^2\Phi = \kappa a^2(\delta p - 2\bar{p}\Phi). \quad (344)$$

If the background is known, this gives us $(1+3+1)=5$ equations for the $(3+3)=6$ unknowns Φ , $\delta\varepsilon$, δp and δU_i . In order to get as many equations as we have unknowns we have to choose an equation of state that allows to express δp in terms of $\delta\varepsilon$.

We want to consider these equations with $\Lambda = 0$ for the special case that the unperturbed and the perturbed spacetime are dust solutions, i.e., $\bar{p} = 0$ and $\delta p = 0$. Then the background spacetime, being a solution to the Friedmann equations for a dust with $k = 0$ and $\Lambda = 0$, must be the Einstein-deSitter universe, recall (174),

$$a = \frac{a_0}{4} \eta^2, \quad t = \frac{a_0 \eta^3}{12c}, \quad (345)$$

$$\mathcal{H}(\eta) = \frac{1}{a(\eta)} \frac{da(\eta)}{d\eta} = \frac{2}{\eta}, \quad (346)$$

$$\frac{d\mathcal{H}}{d\eta} = -\frac{2}{\eta^2}, \quad (347)$$

$$\bar{\varepsilon} = \frac{3a_0}{\kappa a^3} = \frac{3 \times 4^3}{\kappa a_0^2 \eta^6}. \quad (348)$$

As a consequence, (342), (343) and (344) read

$$2\Delta\Phi - \frac{12}{\eta} \frac{\partial\Phi}{\partial\eta} = \frac{12}{\eta^2} \left(\frac{\delta\varepsilon}{\varepsilon} + 2\Phi \right), \quad (349)$$

$$\partial_i \left(\frac{\partial\Phi}{\partial\eta} + \frac{2\Phi}{\eta} \right) = \frac{24}{a_0 c \eta^4} \delta U_i, \quad (350)$$

$$\frac{\partial^2\Phi}{\partial\eta^2} + \frac{6}{\eta} \frac{\partial\Phi}{\partial\eta} = 0. \quad (351)$$

The last equation is a differential equation for Φ alone. After multiplication with η^6 it can be directly integrated,

$$\frac{\partial}{\partial\eta} \left(\eta^6 \frac{\partial\Phi}{\partial\eta} \right) = 0, \quad \eta^6 \frac{\partial\Phi}{\partial\eta} = -5\phi_1(\vec{x}), \quad (352)$$

$$\Phi(\eta, \vec{x}) = \frac{\phi_1(\vec{x})}{\eta^5} + \phi_2(\vec{x}), \quad (353)$$

where ϕ_1 and ϕ_2 are arbitrary functions of $\vec{x} = (x^1, x^2, x^3)$. $\delta\varepsilon$ and δU_i are then given by (349) and (350),

$$\frac{\delta\varepsilon}{\varepsilon} = \frac{\Delta\phi_1}{6\eta^3} + \frac{3\phi_1}{\eta^5} + \frac{\Delta\phi_2}{6}\eta^2 - 2\phi_2, \quad (354)$$

$$\delta U_i = \frac{c a_0}{4} \left(-\frac{\partial_i \phi_1}{2\eta^2} + \frac{\eta^4}{3} \partial_i \phi_2 \right). \quad (355)$$

With (353), (354) and (355) we have found the general solution to the system of equations (349), (350) and (349).

If we express, with the help of (345), the conformal time coordinate η in terms of t , (353) reads

$$\Phi(\eta, \vec{x}) = \phi_1(\vec{x}) \left(\frac{a_0}{12 c t} \right)^{5/3} + \phi_2(\vec{x}). \quad (356)$$

The first term falls off in the course of time. If we wait sufficiently long, the perturbation is given, to within a good approximation, by the second term which is time-independent, i.e., the perturbation is “frozen”. The fact that in a dust universe scalar perturbations of the gravitational field become time-independent for late times was considered to be crucial at a time when people believed that we live in a dust universe without a cosmological constant. Now we believe that there is a cosmological constant that will become dominating for late times. Then the statement that scalar perturbations become “frozen” is no longer true.

For a further discussion of our solution we perform a Fourier expansion of the functions ϕ_1 and ϕ_2 . As for our linearised equations different modes do not couple, we consider just a single mode with wave covector \vec{k} ,

$$\phi_1(\vec{x}) = C_1 e^{i\vec{k}\cdot\vec{x}}, \quad \phi_2(\vec{x}) = C_2 e^{i\vec{k}\cdot\vec{x}}. \quad (357)$$

Note that \vec{k} is dimensionless because \vec{x} is dimensionless. For this mode, the relative density perturbation (354) yields

$$\frac{\delta\varepsilon}{\bar{\varepsilon}} = \left(-\frac{C_1}{6} \frac{k^2}{\eta^3} + \frac{3C_1}{\eta^5} - \frac{C_2}{6} k^2 \eta^2 - 2C_2 \right) e^{i\vec{k}\cdot\vec{x}}, \quad (358)$$

where $k = |\vec{k}|$.

A typical length scale on which the perturbation varies is

$$\ell_p = \frac{a}{k}. \quad (359)$$

On the other hand, the background determines the Hubble length

$$\ell_H = \frac{c}{H} = \frac{c a}{da/dt} = \frac{c a}{da/d\eta} \frac{dt}{d\eta} = \frac{c}{\mathcal{H}} \frac{a}{c} = \frac{c \eta}{2} \frac{a}{c} = \frac{a \eta}{2}. \quad (360)$$

Hence

$$\frac{\ell_H}{\ell_p} = \frac{\eta a k}{2 a} = \frac{\eta k}{2}. \quad (361)$$

For perturbations that vary on length scales that are small in comparison to the Hubble length, $\ell_H \gg \ell_p$, we have $\eta k \gg 1$ and (358) may be approximated as

$$\frac{\delta\varepsilon}{\bar{\varepsilon}} \approx \left(-\frac{C_1}{6} \frac{k^2}{\eta^3} - \frac{C_2}{6} k^2 \eta^2 \right) e^{i\vec{k}\cdot\vec{x}} = \left(-\frac{C_1 k^2 a_0}{72 c t} - \frac{C_2 k^2}{6} \left(\frac{12 c t}{a_0} \right)^{2/3} \right) e^{i\vec{k}\cdot\vec{x}} \quad (362)$$

where we have used (345), i. e., in this case the relative density perturbation has a component that falls off with t^{-1} and a component that increases with $t^{2/3}$.

By contrast, for perturbations that vary on length scales that are big in comparison to the Hubble length, $\ell_H \ll \ell_p$, we have $\eta k \ll 1$ and (358) may be approximated as

$$\frac{\delta\varepsilon}{\bar{\varepsilon}} \approx \left(\frac{3 C_1}{\eta^5} - 2 C_2 \right) e^{i\vec{k}\cdot\vec{x}} = \left(3 C_1 \left(\frac{a_0}{12 c t} \right)^{5/3} - 2 C_2 \right) e^{i\vec{k}\cdot\vec{x}}, \quad (363)$$

i.e., the relative density perturbation has a component that falls off with $t^{-5/3}$ and a component that is time-independent.

We summarise our findings for scalar perturbations of dust solutions in the following way. According to (356), there is a component of the perturbation of the gravitational field that falls off with $t^{-5/3}$ (index 1) and a component that is time-independent (index 2). The corresponding relative density perturbations behave in the following way. For perturbations that vary on small scales, $\ell_H \gg \ell_p$, the first component falls off with t^{-1} whereas the second component grows with $t^{2/3}$. For perturbations that vary on large scales, $\ell_H \ll \ell_p$, the first component falls off with $t^{-5/3}$ whereas the second component is time-independent, just as the perturbation of the gravitational field.

As an application, we calculate the influence of such scalar perturbations on the redshift of light signals and, thereby, on the cosmic background radiation. As a prerequisite, we need the redshift formula in an arbitrary general-relativistic spacetime.

Consider an emitter whose worldline is parametrised by proper time τ_e and an observer whose worldline is parametrised by proper time τ_o , see Figure 43. Denote the four-velocities (i.e., the tangent vector fields to the worldlines) by U_e and U_o , respectively. If two light rays are emitted at times τ_e and $\tau_e + \Delta\tau_e$, they are received at time τ_o and $\tau_o + \Delta\tau_o$ and the frequency ratio is

$$1 + z = \lim_{\Delta\tau_e \rightarrow 0} \frac{\Delta\tau_o}{\Delta\tau_e} = \frac{d\tau_o}{d\tau_e}. \quad (364)$$

The general redshift formula says that

$$1 + z = \frac{g_{\mu\nu} \dot{x}^\nu(s_e) U_e^\mu}{g_{\rho\sigma} \dot{x}^\sigma(s_o) U_o^\rho}. \quad (365)$$

Here $x^\mu(s)$ is the light ray that starts at parameter value $s = s_e$ at the emitter and arrives at the parameter value $s = s_o$ at the observer.

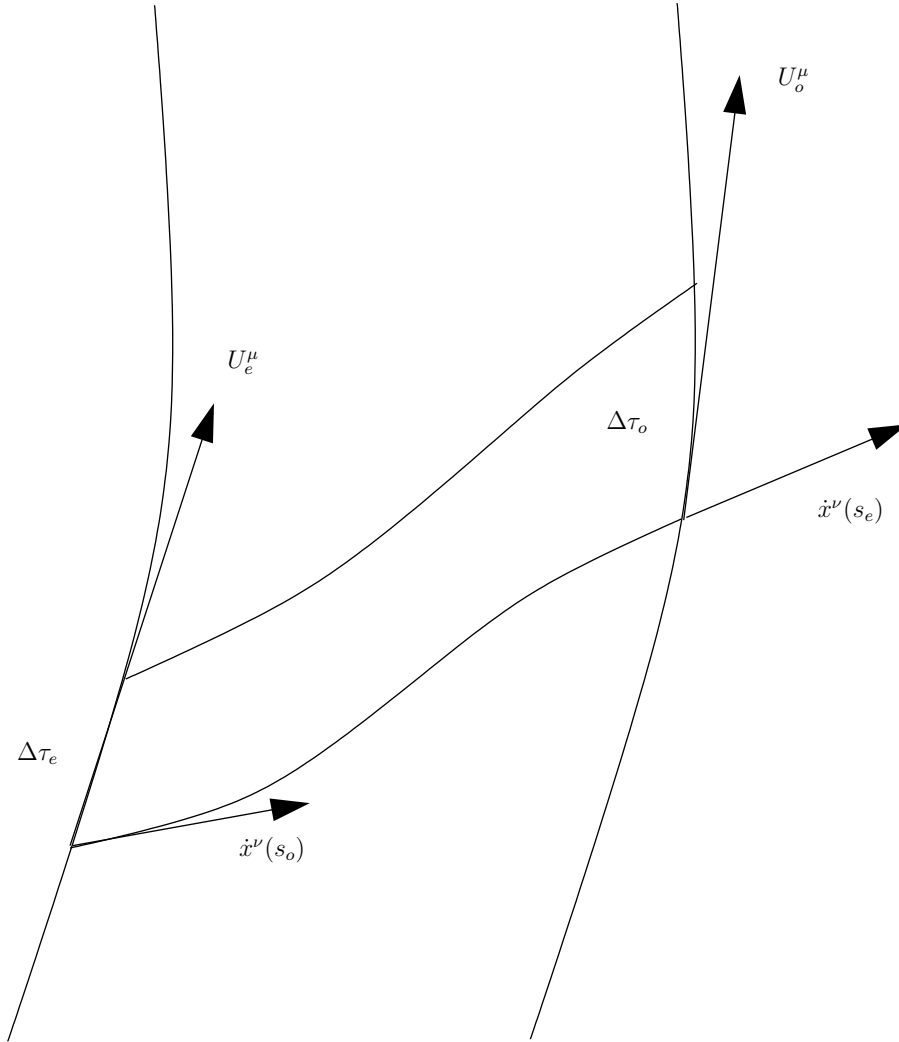


Figure 43: Illustration of the general redshift law

Proof of the general redshift formula:

The following proof is borrowed from D. Brill. It can be found, e.g., in N. Straumann's book [N. Straumann: "General Relativity and Relativistic Astrophysics" Springer (1984)]. We consider the two-surface (possibly with self-intersections) spanned by the light rays from the emitter to the receiver. This two-surface can be labelled by two parameters, s and τ . We choose s as the affine parameter along each light ray, and τ in such a way that it coincides with proper time τ_e on the emitter worldline. Because of the redshift, τ will then *not* coincide with proper time τ_o on the observer worldline. We calculate

$$\partial_s \left(g(\partial_s, \partial_\tau) \right) = g(\nabla_{\partial_s} \partial_s, \partial_\tau) + g(\partial_s, \nabla_{\partial_s} \partial_\tau). \quad (366)$$

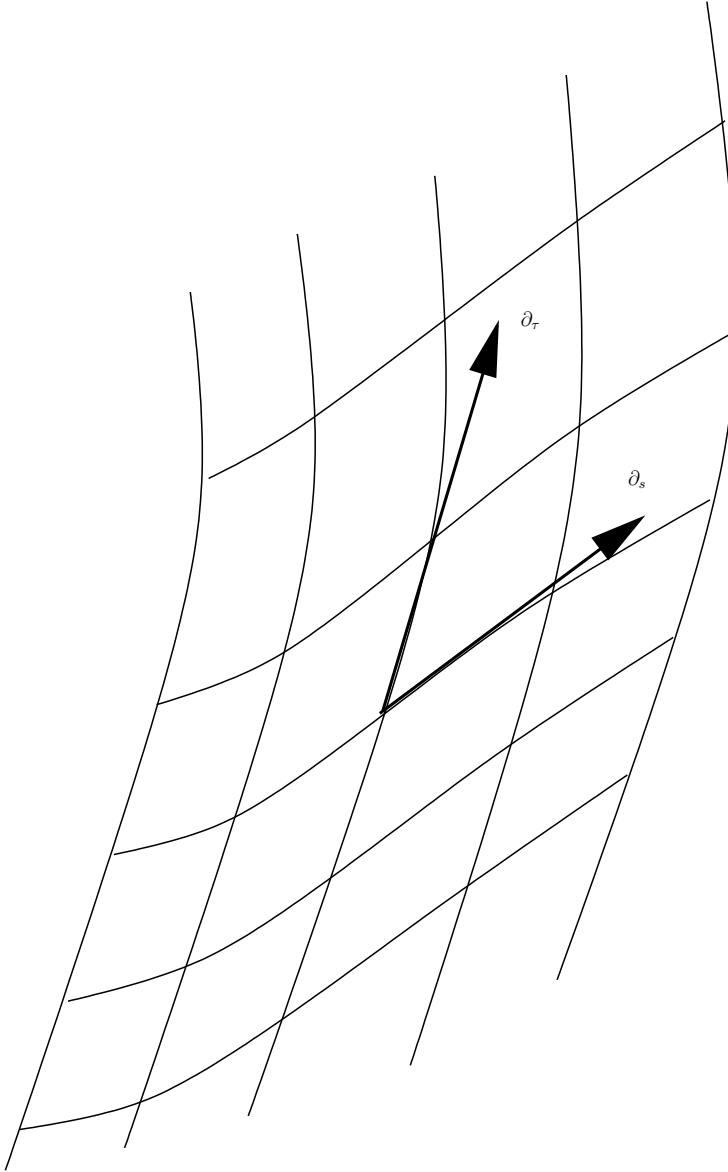


Figure 44: Illustration of the proof of the general redshift law

The first term vanishes, because the light rays are geodesics

$$\nabla_{\partial_s} \partial_s = 0. \quad (367)$$

The second term can be rewritten with the help of the fact that the Levi-Civita connection ∇ is torsion-free, $\nabla_{\partial_s} \partial_\tau = \nabla_{\partial_\tau} \partial_s$. This results in

$$\partial_s \left(g(\partial_s, \partial_\tau) \right) = g(\partial_s, \nabla_{\partial_\tau} \partial_s) = \frac{1}{2} \partial_\tau \left(g(\partial_s, \partial_s) \right) = 0 \quad (368)$$

because the light rays are lightlike, $g(\partial_s, \partial_s) = 0$. We have thus found that

$$g(\partial_s, \partial_\tau) \big|_{s_e} = g(\partial_s, \partial_\tau) \big|_{s_o}. \quad (369)$$

Switching to coordinate notation, this equation reads

$$g_{\mu\nu} \dot{x}^\mu(s_e) U_e^\nu = g_{\rho\sigma} \dot{x}^\rho(s_o) U_o^\sigma \frac{d\tau_o}{d\tau_e} \quad (370)$$

which gives the redshift formula.

We want to evaluate this formula for our perturbed dust universe

$$g_{\mu\nu} dx^\mu dx^\nu = a^2 \left(- (1 + 2\Phi) d\eta^2 + (1 - 2\Phi) \delta_{ij} dx^i dx^j \right) \quad (371)$$

with the Bardeen potential Φ given by (353). The worldlines of observer and emitter are supposed to be integral curves of the four-velocity vector field

$$U^\mu = g^{\mu\rho} U_\rho$$

where U_ρ can be read from (328), hence

$$U^\mu = -g^{\mu\eta} a c (1 + \Phi) + g^{\mu i} \delta U_i. \quad (372)$$

δU_i was calculated for our perturbed dust solution in (355).

The lightlike geodesic we have to consider may be written in the form

$$\eta(s) = \bar{\eta}(s) + \delta\eta(s), \quad x^i(s) = \bar{x}^i(s) + \delta x^i(s) \quad (373)$$

where the overlined quantities are the coordinates of a lightlike geodesic in the unperturbed background spacetime and s is an affine parameter. The Lagrangian for the geodesics is

$$\mathcal{L}(x, \dot{x}) = \frac{1}{2} g_{\mu\nu} \dot{x}^\mu \dot{x}^\nu = \frac{a^2}{2} \left(- (1 + 2\Phi) \dot{\eta}^2 + (1 - 2\Phi) \delta_{ij} \dot{x}^i \dot{x}^j \right). \quad (374)$$

For lightlike geodesics we must have $\mathcal{L} = 0$, hence

$$- (1 + 2\Phi) (\dot{\eta} + \delta\dot{\eta})^2 + (1 - 2\Phi) \delta_{ij} (\dot{\bar{x}}^i + \delta\dot{x}^i) (\dot{\bar{x}}^j + \delta\dot{x}^j) = 0. \quad (375)$$

This gives to zeroth order

$$-\dot{\bar{\eta}}^2 + \delta_{ij} \dot{\bar{x}}^i \dot{\bar{x}}^j = 0 \quad (376)$$

and to first order

$$\begin{aligned} -\dot{\bar{\eta}} \delta \dot{\eta} - \Phi \dot{\bar{\eta}}^2 + \delta_{ij} \dot{\bar{x}}^i \delta \dot{x}^j - \Phi \delta_{ij} \dot{\bar{x}}^i \dot{\bar{x}}^j &= 0, \\ -\dot{\bar{\eta}} \delta \dot{\eta} + \delta_{ij} \dot{\bar{x}}^i \delta \dot{x}^j - 2\Phi \dot{\bar{\eta}}^2 &= 0. \end{aligned} \quad (377)$$

The Euler-Lagrange equation

$$\frac{d}{ds} \left(\frac{\partial \mathcal{L}}{\partial \dot{\bar{\eta}}} \right) = \frac{\partial \mathcal{L}}{\partial \bar{\eta}} \quad (378)$$

reads

$$\frac{d}{ds} \left(-a^2 (1 + 2\Phi) (\dot{\bar{\eta}} + \delta \dot{\eta}) \right) = \frac{a^2}{2} \left(-2\partial_\eta \Phi \dot{\bar{\eta}}^2 - 2\partial_\eta \Phi \delta_{ij} \dot{\bar{x}}^i \dot{\bar{x}}^j + \dots \right) \quad (379)$$

where we have used that $\mathcal{L} = 0$. To zeroth order, this gives

$$\frac{d}{ds} \left(a^2 \dot{\bar{\eta}} \right) = 0 \quad (380)$$

and to first order we find

$$\begin{aligned} \frac{d}{ds} \left(a^2 (\delta \dot{\eta} + 2\Phi \dot{\bar{\eta}}) \right) &= a^2 \partial_\eta \Phi \left(\dot{\bar{\eta}}^2 + \delta_{ij} \dot{\bar{x}}^i \dot{\bar{x}}^j \right), \\ \frac{d}{ds} \left(a^2 \delta \dot{\eta} \right) + 2a^2 \dot{\bar{\eta}} \left(\partial_\eta \Phi \dot{\bar{\eta}} + \partial_i \Phi \dot{\bar{x}}^i \right) &= 2a^2 \partial_\eta \Phi \dot{\bar{\eta}}^2, \\ \frac{d}{ds} \left(a^2 \delta \dot{\eta} \right) &= -2a^2 \dot{\bar{\eta}} \dot{\bar{x}}^i \partial_i \Phi. \end{aligned} \quad (381)$$

Here we have used (376) and (380).

Finally, we have to evaluate the Euler-Lagrange equation

$$\frac{d}{ds} \left(\frac{\partial \mathcal{L}}{\partial \dot{\bar{x}}^i} \right) = \frac{\partial \mathcal{L}}{\partial \bar{x}^i} \quad (382)$$

for $i = 1, 2, 3$ which reads

$$\frac{d}{ds} \left(a^2 (1 - 2\Phi) \delta_{ij} (\dot{\bar{x}}^j + \delta \dot{x}^j) \right) = \frac{a^2}{2} \left(-2\partial_i \Phi \dot{\bar{\eta}}^2 - 2\partial_i \Phi \delta_{kj} \dot{\bar{x}}^k \dot{\bar{x}}^j + \dots \right). \quad (383)$$

To zeroth order this gives

$$\frac{d}{ds} \left(a^2 \delta_{ij} \dot{\bar{x}}^j \right) = 0. \quad (384)$$

To first order we find

$$\begin{aligned}
\frac{d}{ds} \left(a^2 \delta_{ij} (\delta \dot{x}^j - 2 \Phi \dot{\bar{x}}^j) \right) &= -a^2 \partial_i \Phi (\dot{\bar{\eta}}^2 + \delta_{kj} \dot{\bar{x}}^k \dot{\bar{x}}^j), \\
\frac{d}{ds} \left(a^2 \delta_{ij} \delta \dot{x}^j \right) - 2 a^2 \delta_{ij} \dot{\bar{x}}^j (\partial_\eta \Phi \dot{\bar{\eta}} + \partial_k \Phi \dot{\bar{x}}^k) &= -2 a^2 \partial_i \Phi \dot{\bar{\eta}}^2, \\
\frac{d}{ds} \left(a^2 \delta_{ij} \delta \dot{x}^j \right) &= 2 a^2 \left(\delta_{ij} \dot{\bar{x}}^j (\partial_\eta \Phi \dot{\bar{\eta}} + \partial_k \Phi \dot{\bar{x}}^k) - \partial_i \Phi \dot{\bar{\eta}}^2 \right). \tag{385}
\end{aligned}$$

The equations (376), (380) and (384) determine the lightlike geodesics in the background space-time whereas the equations (377), (381) and (385) determine their perturbations. If we introduce the spatial direction vector of the background geodesic,

$$k_i = \frac{\dot{\bar{x}}^i}{\dot{\bar{\eta}}}, \tag{386}$$

these six equations can be rewritten as

$$\delta_{ij} k^i k^j = 1, \tag{387}$$

$$\frac{d}{ds} \left(a^2 \dot{\bar{\eta}} \right) = 0 \tag{388}$$

$$\frac{dk^i}{ds} = 0, \tag{389}$$

$$\delta \dot{\eta} = k_j \delta \dot{x}^j - 2 \Phi \dot{\bar{\eta}}, \tag{390}$$

$$\frac{d}{ds} \left(a^2 \delta \dot{\eta} \right) = -2 a^2 \dot{\bar{\eta}}^2 k^i \partial_i \Phi, \tag{391}$$

$$\frac{d}{ds} \left(a^2 \delta_{ij} \delta \dot{x}^j \right) = 2 a^2 \dot{\bar{\eta}} \left(k_i (\partial_\eta \Phi + k^j \partial_j \Phi) - \partial_i \Phi \dot{\bar{\eta}} \right). \tag{392}$$

For determining the redshift we need to know how $g_{\rho\sigma} \dot{x}^\rho U^\sigma$ changes along the light ray. To that end we calculate

$$\frac{d}{ds} \left(a g_{\rho\sigma} \dot{x}^\rho U^\sigma \right) = \frac{d}{ds} \left(a \dot{x}^\rho U_\rho \right) = \frac{d}{ds} \left(a \dot{\eta} U_\eta + a \dot{x}^i U_i \right).$$

With U_μ from (372) we find

$$\begin{aligned}
\frac{d}{ds} \left(a g_{\rho\sigma} \dot{x}^\rho U^\sigma \right) &= \frac{d}{ds} \left(-a^2 c \dot{\eta} (1 + \Phi) + a \dot{x}^i \delta U_i \right) \\
&= \frac{d}{ds} \left(-a^2 c (\dot{\eta} + \delta \dot{\eta}) (1 + \Phi) + a \dot{x}^i \delta U_i \right) \\
&= \frac{d}{ds} \left(-a^2 c \dot{\eta} - a^2 c \dot{\eta} \Phi - a^2 c \delta \dot{\eta} + a \dot{x}^i \delta U_i \right).
\end{aligned}$$

With (388) this results in

$$\frac{d}{ds} \left(a g_{\rho\sigma} \dot{x}^\rho U^\sigma \right) = a^2 c \dot{\bar{\eta}} \frac{d}{ds} \left(-1 - \Phi - \frac{\delta\dot{\eta}}{\dot{\bar{\eta}}} + \frac{k^i \delta U_i}{a c} \right).$$

With (391) and, again, (388) we find

$$\frac{d}{ds} \left(a g_{\rho\sigma} \dot{x}^\rho U^\sigma \right) = a^2 c \dot{\bar{\eta}} \left(-\partial_\eta \Phi \dot{\bar{\eta}} - \partial_i \Phi \dot{\bar{x}}^i + 2 \dot{\bar{\eta}} k^i \partial_i \Phi + \frac{d}{ds} \left(\frac{k^i \delta U_i}{a c} \right) \right)$$

and hence

$$\frac{d}{ds} \left(a g_{\rho\sigma} \dot{x}^\rho U^\sigma \right) = a^2 c \dot{\bar{\eta}} \underbrace{\left(-\partial_\eta \Phi \dot{\bar{\eta}} + \dot{\bar{\eta}} k^i \partial_i \Phi + \frac{d}{ds} \left(\frac{k^i \delta U_i}{a c} \right) \right)}_{=: Q(s)}. \quad (393)$$

$Q(s)$ is an explicitly known function of the affine parameter s along each ray: k^i is the constant spatial unit vector along the ray, recall (387) and (389) and $\bar{\eta}$ changes along the ray according to (388). The perturbations Φ and δU_i are known from (353) and (355).

Integration of (393) from an observation event (index o) to an emission event (index e) yields

$$a g_{\rho\sigma} \dot{x}^\rho U^\sigma \Big|_e = a g_{\rho\sigma} \dot{x}^\rho U^\sigma \Big|_o + a^2 c \dot{\bar{\eta}} \Big|_o \int_{s_o}^{s_e} Q(s) ds \quad (394)$$

where we have used again (388). Hence

$$\frac{a(\eta_e) g_{\rho\sigma} \dot{x}^\rho U^\sigma \Big|_e}{a(\eta_o) g_{\rho\sigma} \dot{x}^\rho U^\sigma \Big|_o} = 1 + \frac{a^2 c \dot{\bar{\eta}} \Big|_o}{a(\eta_o) g_{\rho\sigma} \dot{x}^\rho U^\sigma \Big|_o} \int_{s_o}^{s_e} Q(s) ds. \quad (395)$$

As the integral on the right-hand side is of first order, we may truncate the factor in front of it after the zeroth order,

$$\frac{g_{\rho\sigma} \dot{x}^\rho U^\sigma \Big|_e}{g_{\rho\sigma} \dot{x}^\rho U^\sigma \Big|_o} = \frac{a(\eta_o)}{a(\eta_e)} \left(1 + \frac{a^2 c \dot{\bar{\eta}} \Big|_o}{a(\eta_o) \bar{g}_{\rho\sigma} \dot{\bar{x}}^\rho \bar{U}^\sigma \Big|_o} \int_{s_o}^{s_e} Q(s) ds \right). \quad (396)$$

Now we insert on the left-hand side the general redshift formula (365) and on the right-hand side we use (328) which implies that $\bar{g}_{\rho\sigma} \dot{\bar{x}}^\rho \bar{U}^\sigma = -a c \dot{\bar{\eta}}$. This gives the perturbed redshift law

$$1 + z = \frac{a(\eta_o)}{a(\eta_e)} \left(1 - \int_{s_o}^{s_e} Q(s) ds \right). \quad (397)$$

To zeroth order we recover, of course, the familiar redshift law in an unperturbed Robertson-Walker universe. We see that the first-order correction is given by an integral over the unperturbed light ray $(\bar{\eta}(s), \bar{x}^i(s))$. This first-order correction is known as the *integrated Sachs-Wolfe effect*. We have restricted our calculation here to the case of scalar perturbations of a dust universe. Sachs and Wolfe calculated this effect in 1967 for arbitrary (i.e. scalar, vector and tensor) perturbations; they didn't use the Bardeen variables (which didn't exist at this time) and rather worked in a gauge where $\delta U^i = 0$.

The integrated Sachs-Wolfe effect gives the major influence of a perturbation on anisotropies in the cosmic background radiation for small ℓ , i.e., on large angular scales. For larger ℓ one also has to take into account another effect (sometimes called the “non-integrated” Sachs-Wolfe effect) resulting from the fact that, because of the perturbation, the temperature of the cosmic background radiation is non-uniform already when it comes into existence at the hypersurface of last scattering. For large ℓ the cosmic background radiation is also influenced by the Sunyaev-Zel’dovich effect which was briefly mentioned on page 84 and by gravitational lensing.

6 Bianchi models

If we want to go beyond the assumptions of homogeneity and isotropy which are inherent to the Robertson-Walker models, we have two possibilities. The first is to use perturbation theory, which in general yields models without any symmetries, but it has the disadvantage that linearising with respect to the perturbations destroys essential features of Einstein’s field equation which is non-linear. The other is to work with exact solutions that have less symmetries than the Robertson-Walker models. In particular, models that are homogeneous but not isotropic have been extensively studied.

For studying homogeneous cosmological models we need the notion of Killing vector fields. Recall that a vector field $K^\mu \partial_\mu$ is called a *Killing vector field* if it satisfies the *Killing equation*

$$\nabla_\mu K_\nu + \nabla_\nu K_\mu = 0. \quad (398)$$

(In coordinate-free notation the Killing equation can be rewritten as $L_K g = 0$ where $L_K g$ is the *Lie derivative* of the metric with respect to the Killing vector field K .) Killing vector fields describe symmetries of the spacetime: In Worksheet 2 we have shown that, near every point where a Killing field $K^\mu \partial_\mu$ is non-zero, we may find a coordinate system such that $K^\mu = \delta_1^\mu$ and the $g_{\rho\sigma}$ are independent of x^1 .

It is obvious that a linear combination $c_1 K_1 + c_2 K_2$ of two Killing vector fields with *constant* coefficients is again a Killing vector field, and it is not difficult to verify that the Lie bracket $[K_1, K_2]$ of two Killing vector fields is again a Killing vector field. (The Lie bracket of two vector fields is their commutator, where we have to view the vector fields as derivative operators acting on a scalar function, $[K_1, K_2]f = K_1 K_2 f - K_2 K_1 f$.) The set of all Killing vector fields on a pseudo-Riemannian manifold is, thus, a Lie algebra. On an n -dimensional manifold, the maximal dimension of this Lie algebra is $n(n+1)/2$.

By definition, a spacetime (i.e., a 4-dimensional Lorentzian manifold) is spatially homogeneous if it admits an algebra of Killing vector fields with 3-dimensional spacelike orbits. (The orbit of a point is the union of all integral curves of Killing vector fields through this point.) The dimension of this Lie algebra cannot be smaller than 3 and it cannot be bigger than $3(3+1)/2 = 6$. We will consider the case that the dimension is equal to 3. The resulting spacetime models

are known as *Bianchi models*. The name refers to the fact that L. Bianchi had classified in the 1890s all 3-dimensional Lie algebras. If the dimension is 4, the spacetime is called *Locally Rotationally Symmetric* (LRS). The case that the dimension is 5 is impossible, and if it is 6 we have a Robertson-Walker model; the latter are special cases of Bianchi models because their 6-dimensional Lie algebra always admits a 3-dimensional subalgebra of Killing vector fields that generate the 3-dimensional orbits.

We briefly review Bianchi's classification of 3-dimensional Lie algebras. Given a 3-dimensional Lie algebra, we may choose a basis (K_1, K_2, K_3) . The Lie bracket of two basis vectors must then be a linear combination of the basis vectors,

$$[K_i, K_j] = C_{ij}^\ell K_\ell, \quad (399)$$

where the so-called *structure constants* C_{ij}^ℓ are real numbers. (As always, we use the summation convention for latin indices $i, j, \dots = 1, 2, 3$). As the commutator of two operators is antisymmetric,

$$[K_i, K_j] = -[K_j, K_i], \quad (400)$$

and satisfies the *Jacobi identity*

$$[[K_i, K_j], K_\ell] + [[K_j, K_\ell], K_i] + [[K_\ell, K_i], K_j] = 0, \quad (401)$$

the structure constants must satisfy

$$C_{ij} = -C_{ji} \quad (402)$$

and

$$\varepsilon^{ijk} C_{ij}^m C_{km}^\ell = 0 \quad (403)$$

where ε^{ijk} is a totally antisymmetric non-zero tensor. We may fix ε^{ijk} by requiring that in the chosen basis $\varepsilon^{123} = 1$. Then any other totally antisymmetric non-zero tensor is given by multiplying ε^{ijk} with a non-zero factor. The antisymmetry of the structure constants with respect to the lower indices implies that the same information as in the C_{ij}^ℓ is in the second-rank tensor

$$t^{ij} = \varepsilon^{imn} C_{mn}^j. \quad (404)$$

We decompose t^{ij} into symmetric and antisymmetric parts,

$$t^{ij} = n^{ij} + \varepsilon^{ijk} a_k, \quad n^{ij} = n^{ji}. \quad (405)$$

With a bit of algebra one verifies that then the Jacobi identity is satisfied if and only if

$$n^{ij} a_j = 0. \quad (406)$$

One says that a 3-dimensional Lie algebra is of

- Bianchi Class A if $(a_1, a_2, a_3) = (0, 0, 0)$,
- Bianchi Class B if $(a_1, a_2, a_3) \neq (0, 0, 0)$.

A change of the basis,

$$\tilde{K}_i = L_i^j K_j, \quad (407)$$

preserves the condition $\varepsilon^{123} = 1$ if $\det(L_i^j) = 1$. Under such a transformation,

$$\tilde{n}^{ij} = \left(L^{-1}\right)^i_k \left(L^{-1}\right)^j_\ell n^{k\ell}, \quad \tilde{a}_i = L_i^j a_j. \quad (408)$$

We may also change to another totally antisymmetric tensor, $\hat{\varepsilon}^{ijk} = \lambda \varepsilon^{ijk}$. Then

$$\hat{n}^{ij} = \lambda n^{ij}, \quad \hat{a}_j = a_j. \quad (409)$$

In the case of Bianchi Class A, a transformation (408) with an orthogonal matrix (L_i^j) may be used to diagonalise the matrix (n^{ij}) ,

$$(n^{ij}) = \begin{pmatrix} n_1 & 0 & 0 \\ 0 & n_2 & 0 \\ 0 & 0 & n_3 \end{pmatrix}. \quad (410)$$

This may be followed by a transformation (408) with a diagonal matrix (L_i^j) to set all non-zero diagonal elements of (n^{ij}) equal in magnitude. Finally, a transformation (409) may be used to set all the non-zero diagonal elements equal to 1 or -1 . We may choose the sign of λ such that the number of negative diagonal elements is smaller than or equal to the number of positive ones. We are then left with the following Bianchi types within Class A:

n_1	n_2	n_3	Bianchi type
0	0	0	I
1	0	0	II
0	1	-1	VI ₀
0	1	1	VII ₀
1	1	-1	VIII
1	1	1	IX

For Lie algebras of Bianchi Class B we may choose a transformation (408) with (L_i^j) orthogonal such that we achieve the form

$$(n^{ij}) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & n_2 & 0 \\ 0 & 0 & n_3 \end{pmatrix}, \quad (a_i) = \begin{pmatrix} a \\ 0 \\ 0 \end{pmatrix}. \quad (411)$$

Here we have used that n^{ij} is symmetric and that $n^{ij}a_j = 0$. This form is preserved under transformations (408) with

$$(L_i^j) = \begin{pmatrix} \frac{1}{b_2 b_3} & 0 & 0 \\ 0 & b_2 & 0 \\ 0 & 0 & b_3 \end{pmatrix}. \quad (412)$$

If $n_2 = n_3 = 0$, we can use such a transformation for setting a equal to 1. If $n_2 = 0$ and $n_3 \neq 0$, we do the same thing and simultaneously transform n_3 to 1 or -1 . We may then use a transformation (T2) for setting n_3 equal to 1. The situation is more difficult if $n_2 n_3 \neq 0$. Then we see that the remaining transformations (408) leave

$$h := \frac{a^2}{n_2 n_3} \quad (413)$$

invariant. We may choose such a transformation (408) for setting n_2 and n_3 equal to 1 or -1 , and we may use a transformation (409) for transforming the case that both are negative to the case that both are positive. Then, however, there is no further freedom for normalising a ; the resulting Bianchi class will depend on the parameter h that may take all real values (non-zero for Bianchi Class B). This gives us the following Bianchi types for Class B.

a	n_1	n_2	n_3	Bianchi type
1	0	0	0	V
1	0	0	1	IV
$\sqrt{-h}$	0	1	-1	VI _h ($h < 0$)
\sqrt{h}	0	1	1	VII _h ($h > 0$)

Bianchi type III is missing in the table because it is the same as VI₋₁.

It is our goal to study Bianchi models of type I in some detail, first for vacuum and then for dust. Bianchi I is the simplest type; all the structure constants are zero, i.e., if we choose a basis (K_1, K_2, K_3) of Killing vector fields that generate the Bianchi symmetry, we have

$$[K_i, K_j] = 0. \quad (414)$$

Note that K_1 , K_2 and K_3 must be linearly independent at each point because they are assumed to generate 3-dimensional spacelike hypersurfaces. Then the condition that the Lie bracket vanishes implies that we can choose, on each of these 3-dimensional hypersurfaces, coordinates (x, y, z) such that

$$K_1 = \partial_x, \quad K_2 = \partial_y, \quad K_3 = \partial_z. \quad (415)$$

As the fourth coordinate, we choose proper time t along the timelike curves perpendicular to the homogeneous slices. As ∂_x , ∂_y and ∂_z are Killing vector fields, the metric coefficients are functions of t only. For one time t , we may choose the spatial coordinate axes such that $g_{ij} = 0$ for $i \neq j$. We try to find solutions to the field equations, first for vacuum and then for dust, such that the metric remains diagonal for all times, i.e., we assume that the metric is of the form

$$g = -c^2 dt^2 + X(t)^2 dx^2 + Y(t)^2 dy^2 + Z(t)^2 dz^2. \quad (416)$$

In a Robertson-Walker universe we had one scale factor $a(t)$, now we have three scale factors $X(t)$, $Y(t)$, $Z(t)$. Correspondingly, there are three Hubble parameters which we denote

$$A(t) = \frac{X'(t)}{X(t)}, \quad B(t) = \frac{Y'(t)}{Y(t)}, \quad Z(t) = \frac{Z'(t)}{Z(t)}. \quad (417)$$

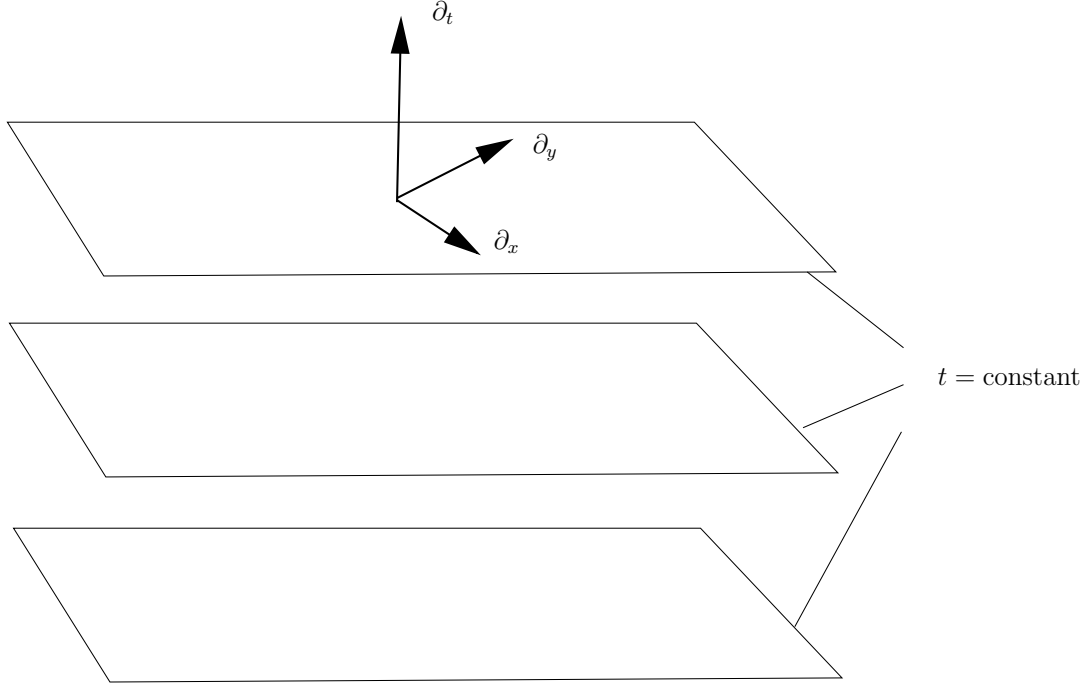


Figure 45: Bianchi model

We also use the abbreviation

$$\theta(t) = A(t) + B(t) + C(t). \quad (418)$$

For the Ricci tensor of our Bianchi I metric we find

$$R_{tt} = -\theta' - A^2 - B^2 - C^2, \quad (419)$$

$$R_{xx} = \frac{X^2}{c^2} (A' + \theta A), \quad (420)$$

$$R_{yy} = \frac{Y^2}{c^2} (B' + \theta B), \quad (421)$$

$$R_{zz} = \frac{Z^2}{c^2} (C' + \theta C). \quad (422)$$

The off-diagonal elements vanish. The Ricci scalar reads

$$R = \frac{2}{c^2} (\theta' + A^2 + B^2 + C^2 + AB + BC + CA). \quad (423)$$

We first determine the general solution to the vacuum field equation without a cosmological constant,

$$R_{\mu\nu} = 0. \quad (424)$$

Then we must have

$$0 = R_{tt} + \frac{c^2}{2} R = AB + BC + CA. \quad (425)$$

This implies that

$$\theta^2 = (A + B + C)^2 = A^2 + B^2 + C^2. \quad (426)$$

Inserting this result into the equation $R_{tt} = 0$ yields a differential equation for θ ,

$$\theta' + \theta^2 = 0. \quad (427)$$

For solving this differential equation we first consider the case that $\theta = 0$ on a t -interval. Then equation (426) implies that $A = B = C = 0$ on this interval, i.e., that X , Y and Z are constants. This gives Minkowski spacetime which is already known to us as a solution to the vacuum field equation. We may thus restrict in the following to the case that $\theta \neq 0$ on a t -interval. This includes all cases where θ has isolated zeros which would occur at the boundary of the considered interval. If $\theta \neq 0$, the differential equation (427) can be solved by separation of variables,

$$\frac{d\theta}{\theta^2} = -dt, \quad -\frac{1}{\theta} = -(t - t_i). \quad (428)$$

As we are free to choose the origin of the t coordinate where we like, we choose the integration constant t_i equal to 0, i.e.,

$$\theta(t) = \frac{1}{t}. \quad (429)$$

With this result at hand, we can evaluate the equation $R_{xx} = 0$ which yields

$$\frac{dA}{dt} + \frac{A}{t} = 0, \quad \frac{dA}{A} = -\frac{dt}{t}, \quad (430)$$

$$\ln A = -\ln t + \ln p \quad (431)$$

with an integration constant p , hence

$$A(t) = \frac{p}{t}. \quad (432)$$

With the definition of A this may be rewritten as

$$\frac{dX}{X} = p \frac{dt}{t}, \quad \ln X = p \ln t - p \ln t_0 \quad (433)$$

with an integration constant t_0 , hence

$$X(t) = \left(\frac{t}{t_0}\right)^p. \quad (434)$$

Similarly, evaluation of the equations $R_{yy} = 0$ and $R_{zz} = 0$ yields

$$Y(t) = \left(\frac{t}{t_0}\right)^q, \quad Z(t) = \left(\frac{t}{t_0}\right)^r, \quad (435)$$

with integration constants q and r . Note that we could choose the same integration constant t_0 for all three components because we are free to shift the origin of the spatial coordinate system.

The constants p , q and r are not independent of each other: Condition (429) requires

$$p + q + r = 1, \quad (436)$$

and condition (426) requires

$$p^2 + q^2 + r^2 = 1. \quad (437)$$

In (p, q, r) space, (436) determines a plane and (437) determines a sphere, so the values of (p, q, r) are restricted to a circle which is known as the *Kasner circle*. It is the unique circle that passes through the three points $(1, 0, 0)$, $(0, 1, 0)$ and $(0, 0, 1)$, see Figure 46. We have thus found that the vacuum solutions of Bianchi type I, with the exception of the Minkowski spacetime, are all metrics of the form

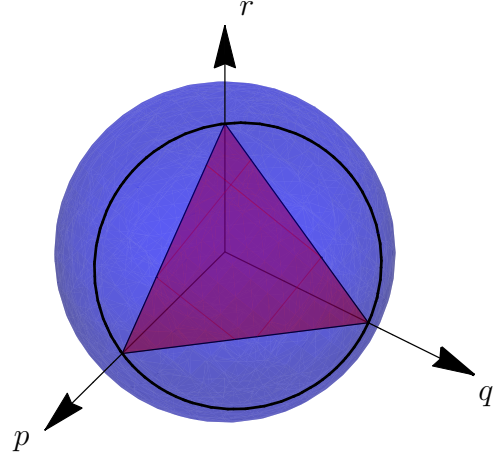


Figure 46: Kasner circle

$$g = -c^2 dt^2 + \left(\frac{t}{t_0}\right)^{2p} dx^2 + \left(\frac{t}{t_0}\right)^{2q} dy^2 + \left(\frac{t}{t_0}\right)^{2r} dz^2 \quad (438)$$

where (p, q, r) is any point on the Kasner circle. These vacuum solutions were found in 1921 by the US American mathematician E. Kasner who did not refer to the Bianchi classification. Therefore, (438) is known as the *Kasner metric*. Depending on the signs of the coefficients p , q and r , a Kasner universe is expanding in some spatial directions and contracting in others. The signs of (p, q, r) are restricted to the following two cases, as can be read from Figure 46.

- Two Kasner coefficients vanish and the third one is equal to 1. This is true at the vertices of the triangle in Figure 46. We consider the case that $p = 1$, $q = r = 0$. Then the metric reads

$$g = -c^2 dt^2 + \left(\frac{t}{t_0}\right)^2 dx^2 + dy^2 + dz^2. \quad (439)$$

A coordinate transformation

$$\tilde{t} = t \cosh \frac{x}{ct_0}, \quad \tilde{x} = ct \sinh \frac{x}{ct_0}, \quad \tilde{y} = y, \quad \tilde{z} = z, \quad (440)$$

reveals that this is Minkowski spacetime,

$$g = -c^2 d\tilde{t}^2 + d\tilde{x}^2 + d\tilde{y}^2 + d\tilde{z}^2. \quad (441)$$

The original coordinates are known as *Rindler coordinates*. They cover the “Rindler wedge” $\tilde{x} > c|\tilde{t}|$ and the t -lines are the worldlines of observers with constant acceleration (“Rindler observers”). An analogous result holds, of course, for $(p, q, r) = (0, 1, 0)$ and $(p, q, r) = (0, 0, 1)$.

- All three Kasner coefficients are different from zero, where two of them are positive and the third one is negative. This is true at all points on the Kasner circle except at the vertices of the triangle, see again Figure 46. We consider the case that $p > 0$, $q > 0$ and $r < 0$. Then (438) gives a universe with a singularity. If we consider the time interval

$0 < t < \infty$, it is an initial singularity. This may be called a “big bang”, but in contrast to the singularity in a Robertson-Walker universe the Kasner singularity is anisotropic. If it is approached backwards in time, only the x and y components of a comoving volume element shrink to zero while the z component blows up. This is called a *cigar singularity*. Of course, an analogous result holds in the cases that $p < 0$ and $q < 0$, respectively, which follows just by a permutation of coordinates.

We summarise our findings on the Bianchi I vacuum solutions in the following way: Generically, we have a cigar singularity. There are exceptional cases where the solution is the Minkowski spacetime.

We will now discuss the Bianchi I universes with a dust source. It is our main goal to study the effect of the anisotropy on the initial singularity. As the cosmological constant is relevant only for the late universe, we may restrict ourselves to the case $\Lambda = 0$, i.e., to the field equation

$$R_{\rho\sigma} - \frac{R}{2} g_{\rho\sigma} = \kappa T_{\rho\sigma} \quad (442)$$

where

$$T_{\rho\sigma} = \mu U_\rho U_\sigma. \quad (443)$$

We will assume that the four-velocity of the dust is perpendicular to the homogeneous slices, i.e.,

$$U^\rho = \delta_t^\rho, \quad U_\sigma = g_{\sigma\rho} U^\rho = g_{\sigma t} = -c^2 \delta_\sigma^t. \quad (444)$$

With the components of the Ricci tensor and the Ricci scalar given by (419), (420), (421), (422) and (423), the (tt) , (xx) , (yy) and (zz) components of the field equation read

$$AB + BC + CA = \kappa c^4 \mu, \quad (445)$$

$$A' + \theta A - \theta' - \theta^2 + AB + BC + CA = 0, \quad (446)$$

$$B' + \theta B - \theta' - \theta^2 + AB + BC + CA = 0, \quad (447)$$

$$C' + \theta C - \theta' - \theta^2 + AB + BC + CA = 0. \quad (448)$$

The off-diagonal components of the field equation reduce to the triviality $0 = 0$. Following the same strategy as in the vacuum case, we first derive a differential equation for θ alone, then we solve for the other unknown quantities.

Adding (446), (447) and (448) together yields

$$\theta' + \theta^2 - 3\theta' - 3\theta^2 + 3(AB + BC + CA) = 0, \quad (449)$$

$$2\theta' + 2\theta^2 = 3(AB + BC + CA). \quad (450)$$

Differentiating with respect to t results in

$$\begin{aligned} (\theta' + \theta^2)' &= \frac{3}{2} (A'B + AB' + B'C + BC' + C'A + CA') \\ &= \frac{3}{2} ((B+C)A' + (C+A)B' + (A+B)C'). \end{aligned} \quad (451)$$

With the help of (446), (447) and (448) this can be rewritten as

$$\begin{aligned}
(\theta' + \theta^2)' &= \frac{3}{2} \left((B+C) (-\theta A + \theta' + \theta^2 - AB - BC - CA) \right. \\
&\quad + (C+A) (-\theta B + \theta' + \theta^2 - AB - BC - CA) \\
&\quad \left. + (A+B) (-\theta C + \theta' + \theta^2 - AB - BC - CA) \right) \\
&= \frac{3}{2} \left(-2\theta (AB + BC + CA) + 2(\theta' + \theta^2 - AB - BC - CA)(A+B+C) \right) \\
&= -3\theta (AB + BC + CA) + 3(\theta' + \theta^2 - AB - BC - CA)\theta \\
&= -6\theta (AB + BC + CA) + 3(\theta' + \theta^2)\theta
\end{aligned} \tag{452}$$

Inserting (450) yields

$$(\theta' + \theta^2)' = -4\theta(\theta' + \theta^2) + 3(\theta' + \theta^2)\theta,$$

$$\theta'' + 2\theta\theta' = -\theta\theta' - \theta^3,$$

$$\theta'' + 3\theta\theta' + \theta^3 = 0. \tag{453}$$

We have thus achieved our goal of deriving a differential equation for θ alone. With the ansatz

$$\theta = \frac{v'}{v}, \quad \theta' = \frac{v''}{v} - \frac{v'^2}{v^2}, \quad \theta'' = \frac{v'''}{v} - \frac{3v''v'}{v^2} + \frac{2v'^3}{v^3} \tag{454}$$

the differential equation (453) can be rewritten as

$$\frac{v'''}{v} - \cancel{\frac{3v''v'}{v^2}} + \cancel{\frac{2v'^3}{v^3}} + \frac{3v'}{v} \left(\cancel{\frac{v''}{v}} - \cancel{\frac{v'^2}{v^2}} \right) + \cancel{\frac{v'^3}{v^3}} = 0, \tag{455}$$

i.e.,

$$v''' = 0. \tag{456}$$

The general solution is

$$v(t) = \alpha t^2 + \beta t + \gamma \tag{457}$$

with constants α , β and γ , hence

$$\theta(t) = \frac{2\alpha t + \beta}{\alpha t^2 + \beta t + \gamma}. \tag{458}$$

By (445) and (450),

$$\frac{3}{2} \kappa c^4 \mu = \theta' + \theta^2 = \frac{v''}{v} - \frac{v'^2}{v^2} + \frac{v'^2}{v^2} = \frac{2\alpha}{\alpha t^2 + \beta t + \gamma}. \quad (459)$$

We see that $\alpha = 0$ gives the vacuum case $\mu = 0$ which was already covered, so we are only interested in the case that $\alpha \neq 0$. From (458) and (459) we see that then we may assume, without loss of generality, that $\alpha = 1$. Moreover, as we are free to choose the origin of the time coordinate as we like, we may set β equal to zero, hence

$$\theta(t) = \frac{2t}{t^2 + \gamma} \quad (460)$$

and

$$\frac{3}{2} \kappa c^4 \mu(t) = \frac{2}{t^2 + \gamma}. \quad (461)$$

What remains to be done is to determine A , B and C from (446), (447) and (448), respectively. With (445) and (450), equation (446) can be rewritten as

$$0 = A'(t) + \theta(t) A(t) - \frac{1}{2} \kappa c^4 \mu(t) = A'(t) + \frac{2t A(t)}{t^2 + \gamma} - \frac{2}{3(t^2 + \gamma)}, \quad (462)$$

$$(t^2 + \gamma) A'(t) + 2t A(t) - \frac{2}{3} = 0, \quad (463)$$

$$\frac{d}{dt} \left((t^2 + \gamma) A(t) \right) = \frac{2}{3}, \quad (464)$$

$$(t^2 + \gamma) A(t) = \frac{2}{3} (t + \tilde{p}), \quad (465)$$

$$A(t) = \frac{2(t + \tilde{p})}{3(t^2 + \gamma)} \quad (466)$$

with an integration constant \tilde{p} . Analogously we find from (447) and (448) that

$$B(t) = \frac{2(t + \tilde{q})}{3(t^2 + \gamma)}, \quad (467)$$

$$C(t) = \frac{2(t + \tilde{r})}{3(t^2 + \gamma)}. \quad (468)$$

The integration constants \tilde{p} , \tilde{q} and \tilde{r} are not independent. As $A(t) + B(t) + C(t) = \theta(t)$, we must have

$$\frac{\cancel{2} (3t + \tilde{p} + \tilde{q} + \tilde{r})}{3 \cancel{(t^2 + \gamma)}} = \frac{\cancel{2} t}{\cancel{t^2 + \gamma}}, \quad (469)$$

hence

$$\tilde{p} + \tilde{q} + \tilde{r} = 0. \quad (470)$$

Moreover, by (445) we must have

$$\frac{\mathcal{A}(3t^2 + 2(\tilde{p} + \tilde{q} + \tilde{r}) + \tilde{p}\tilde{q} + \tilde{q}\tilde{r} + \tilde{r}\tilde{p})}{3\cancel{t^2}(t^2 + \gamma)^2} = \frac{\mathcal{A}}{\cancel{3}(t^2 + \gamma)},$$

$$3t^2 + 0 + \tilde{p}\tilde{q} + \tilde{q}\tilde{r} + \tilde{r}\tilde{p} = 3(t^2 + \gamma),$$

$$3\gamma = \tilde{p}\tilde{q} + \tilde{q}\tilde{r} + \tilde{r}\tilde{p}, \quad (471)$$

hence

$$0 = (\tilde{p} + \tilde{q} + \tilde{r})^2 = \tilde{p}^2 + \tilde{q}^2 + \tilde{r}^2 + 2(\tilde{p}\tilde{q} + \tilde{q}\tilde{r} + \tilde{r}\tilde{p}), \quad (472)$$

$$\tilde{p}^2 + \tilde{q}^2 + \tilde{r}^2 = -6\gamma. \quad (473)$$

This equation demonstrates that γ cannot be positive. If $\gamma = 0$, we have $A(t) = B(t) = C(t) = 2/(3t)$ which gives the spatially flat Robertson-Walker dust universe without a cosmological constant, i.e., the Einstein-deSitter universe. As we know this case already sufficiently well, we assume in the following that $\gamma < 0$. We may then set

$$\gamma = -t_0^2, \quad t_0 > 0. \quad (474)$$

By (466), the scale factor $X(t)$ is then given by integrating the equation

$$\frac{X'(t)}{X(t)} = A(t) = \frac{2(t + \tilde{p})}{3(t^2 - t_0^2)}, \quad (475)$$

$$\frac{dX}{X} = \frac{2(t + \tilde{p}) dt}{3(t^2 - t_0^2)}, \quad (476)$$

which gives an elementary integral that can be looked up in an integral table,

$$X(t) = (t - t_0)^p (t + t_0)^{\frac{2}{3}-p}, \quad p = \frac{1}{3} \left(1 + \frac{\tilde{p}}{t_0} \right). \quad (477)$$

Analogously,

$$Y(t) = (t - t_0)^q (t + t_0)^{\frac{2}{3}-q}, \quad q = \frac{1}{3} \left(1 + \frac{\tilde{q}}{t_0} \right), \quad (478)$$

$$Z(t) = (t - t_0)^r (t + t_0)^{\frac{2}{3}-r}, \quad r = \frac{1}{3} \left(1 + \frac{\tilde{r}}{t_0} \right). \quad (479)$$

The relations (470) and (473) of the coefficients \tilde{p} , \tilde{q} and \tilde{r} imply that p , q and r satisfy the Kasner relations (436) and (437),

$$p + q + r = \frac{1}{3} \left(3 + \frac{\tilde{p} + \tilde{q} + \tilde{r}}{t_0} \right) = 1 + 0, \quad (480)$$

$$p^2 + q^2 + r^2 = \frac{1}{9} \left(3 + \frac{2(\tilde{p} + \tilde{q} + \tilde{r})}{t_0} + \frac{\tilde{p}^2 + \tilde{q}^2 + \tilde{r}^2}{t_0^2} \right) = \frac{1}{9} \left(3 + 0 + \frac{6t_0^2}{t_0^2} \right) = 1. \quad (481)$$

The metric reads

$$g = -c^2 dt^2 + (t - t_0)^{2p} (t + t_0)^{\frac{4}{3}-2p} dx^2 + (t - t_0)^{2q} (t + t_0)^{\frac{4}{3}-2q} dy^2 + (t - t_0)^{2r} (t + t_0)^{\frac{4}{3}-2r} dz^2. \quad (482)$$

The metric is regular on the interval $-\infty < t < -t_0$, on the interval $-t_0 < t < t_0$ and on the interval $t_0 < t < \infty$. We consider the latter case which is a universe with an initial singularity but no final singularity. The signs of p , q and r determine the behaviour of the scale factors if the singularity is approached. We have already discussed that the Kasner relations can be satisfied only in the following two cases:

(i) One Kasner coefficient is equal to 1 and the other two are zero. This is true at the vertices of the triangle in Figure 46. In the vacuum case the metric was then given by Minkowski spacetime in Rindler coordinates. In the dust case, (482) gives a universe in which, if the initial singularity is approached from the future, the scale factor shrinks to zero in one dimension and it stays finite in the other two dimensions. This is called a *pancake singularity*.

(ii) Two Kasner coefficients are positive and the third one is negative. This is the generic case, i.e., it is true at all points on the Kasner circle except at the vertices of the triangle. As in the vacuum case, we read from the metric that then, if the singularity is approached from the future, the scale factor shrinks to zero in two spatial dimensions and it blows up in the third one. We have already mentioned that this is called a *cigar singularity*.

One might have thought that the initial singularity of Robertson-Walker universes is an artifact of the assumed isotropy. Now we see that this is not true. At least for a Bianchi I dust universe, we have demonstrated that dropping the assumption of isotropy does *not* avoid the formation of a singularity. The only difference in comparison to the Robertson-Walker case is in the fact that the singularity is approached in an anisotropic fashion. Generically, Bianchi I dust universes feature a cigar singularity, just as Bianchi I vacuum universes. We may thus say that, although in a Bianchi I dust universe the density $\mu(t)$ goes to infinity if the singularity is approached, generically the dust has no influence on the character of the singularity.

7 Singularity theorems

When studying Robertson-Walker universes it seems likely that the occurrence of a singularity is an artifact of the high symmetry. This is analogous to the investigation of gravitational collapse where Oppenheimer and Snyder had shown in 1939 that a spherically symmetric ball of dust ends up in a singularity; also in this case, it seemed likely that there is no longer a singularity if spherical symmetry is broken. Our discussion of Bianchi I dust models has given a first indication that singularities might not be an artifact of high symmetries; Bianchi I models are still homogeneous but not isotropic, so one might have expected that they would avoid a singularity. However, we have seen that in the Bianchi I case only the character of the singularity is changed (from a point singularity generically to a cigar singularity), but not the fact that there is a singularity.

During the 1960s it became evident that the occurrence of singularities is a general feature of Einstein's field equation which has nothing to do with symmetries. There were two lines of research to this effect.

- In the Soviet Union, members of the Landau school, in particular V. Belinsky, I. Khalatnikov and E. Lifshitz (BKL), investigated the set of all initial conditions for Einstein's field equation that lead to a singularity. It was a characteristic feature of their work to concentrate on features that are independent of the matter model. (BKL considered cosmological solutions where the singularity is in the past of the initial hypersurface; for gravitational collapse it is in the future.) In an early paper, Lifshitz and Khalatnikov had claimed that almost all initial conditions lead to singularity-free solutions. Later the Russian scientists realised that this was an error and they found heuristic evidence that, on the contrary, singularities are the rule rather than the exception. However, they did not succeed in rigourously proving a theorem to this effect. Nonetheless, their work is very important because it gave some insight on how a singularity is approached.
- In the United Kingdom R. Penrose and S. Hawking proved a series of theorems demonstrating that singularities occur under rather generic conditions. There are four such theorems: The first one by Penrose (1965) is relevant for gravitational collapse, the second and third by Hawking (1967) are relevant for cosmology and the fourth one by Penrose and Hawking together (1970) is relevant for both situations.

In what follows we briefly summarise the content of Hawking's singularity theorems. The proofs are so involved that we will not even touch upon them. Details can be found in the book by S. Hawking and G. Ellis [“The large-scale structure of space-time”, Cambridge University Press (1973)].

It is important to realise that the Penrose-Hawking singularity theorems do not prove the existence of a singularity in the sense that the energy density or a curvature invariant becomes infinite. What is proven is that there are timelike or lightlike geodesics that are incomplete (in the past for cosmological solutions and in the future for gravitational collapse). For timelike geodesics, this means that for a freely falling observer the world ends at a finite proper time which clearly indicates a pathological situation. For lightlike geodesics the affine parameter cannot be interpreted as the reading of a clock, but also incompleteness of a lightlike geodesic seems to be something pathological because a lightlike geodesic is the history of a photon. An example why it is not sufficient to study timelike incompleteness is given by the Reissner-Nordström metric which features a curvature singularity at $r = 0$ where no timelike but lightlike geodesics terminate. Of course, geodesics become incomplete in a trivial way if we remove points from a perfectly regular spacetime. Therefore, we say that a spacetime is singular if it is *inextendible* and contains an incomplete timelike or lightlike geodesic.

Having agreed on the definition of singularities, the task is to formulate hypotheses that are physically reasonable and predict the existence of a singularity. The Penrose-Hawking theorems use three types of hypotheses:

- First one needs a condition on the Ricci tensor which makes sure that gravity is attractive in the sense that it makes the worldlines of freely falling objects converge. In conjunction with Einstein's field equation, such a condition can be re-interpreted as an *energy condition*. For the Hawking singularity theorems the condition

$$R_{\mu\nu}K^\mu K^\nu \geq 0 \quad \text{if} \quad g_{\mu\nu}K^\mu K^\nu \leq 0 \quad (483)$$

is used. In view of the Jacobi equation (also known as the equation of geodesic deviation), this condition means that on averaging over directions gravity is attractive for freely

falling particles and photons. If Einstein's field equation for a perfect fluid without a cosmological constant is assumed, it can be rewritten as

$$\varepsilon + p \geq 0, \quad \varepsilon + 3p \geq 0 \quad (484)$$

and is known as the *strong energy condition*. It is obviously satisfied for a perfect fluid with positive energy density and positive pressure. It is violated, however, for “dark energy”, i.e., for a perfect fluid mimicking a positive cosmological constant where $p = -\varepsilon$ is negative, recall Problem 2 of Worksheet 4.

- Then one needs a condition either on the topology or on the causal structure of spacetime. In one of the two Hawking theorems one considers a “closed universe”, i.e., one assumes a compact spatial topology. In the other Hawking theorem one assumes that there are no closed timelike curves. It is widely accepted that closed timelike curves should be forbidden because they lead to the paradox that one could travel into one's own past and kill one's parents before one is borne. Actually, in the Hawking theorem a slightly stronger assumption is needed which is known as the “strong causality condition”: Every neighbourhood of a point p contains a neighbourhood of p that no timelike or lightlike curve through p intersects more than once. This is a way of saying that it is not only forbidden for a timelike or lightlike curve to come back exactly to p but also to come back arbitrarily close to p .
- Finally, a third assumption is needed that makes sure that, at one time, the spacetime has the tendency to “contract” (towards the past for cosmology and towards the future for gravitational collapse). Such an initial condition is formulated with the help of a vector field $V = V^\mu \partial_\mu$ that satisfies $g(V, V) = -c^2$ and $\nabla_V V = 0$, i.e., its worldlines are timelike geodesics parametrised by proper time. For such a vector field, the scalar field $\theta = \nabla_\mu V^\mu$ is called the *expansion*. It measures if neighbouring worldlines approach each other ($\theta < 0$) or move away from each other ($\theta > 0$). The idea is to prove that, if a contracting initial condition is prescribed, and if the other assumptions of the theorem are satisfied, then the collapse cannot be stopped and will lead to a singularity.

We now give the precise formulation of the two Hawking theorems.

Theorem 1 (Hawking, 1967): Spacetime cannot be timelike and lightlike geodesically complete in the past if the following three assumptions hold:

- (a) The strong energy condition is satisfied,

$$R_{\mu\nu}K^\mu K^\nu \geq 0 \quad \text{if} \quad g_{\mu\nu}K^\mu K^\nu \leq 0.$$

- (b) There is a spacelike compact 3-dimensional submanifold S without boundary.
- (c) Let $V = V^\mu \partial_\mu$ be the past oriented vector field with $g(V, V) = -c^2$ whose integral curves are the timelike geodesics orthogonal to S . Then the expansion $\theta = \nabla_\mu V^\mu$ satisfies $\theta|_S < 0$.

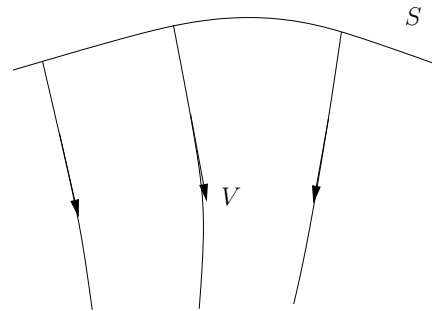


Figure 47: Illustration of Theorem 1, assumption (c)

Note that in the theorem it is not assumed that the spacetime be inextendible. However, if we start with a spacetime where the assumptions (a), (b) and (c) hold, the theorem says that this spacetime cannot be extended to a timelike and lightlike geodesically complete spacetime without violating one of these three assumptions.

We have good observational evidence that we live in a universe that is expanding. If our universe is spatially compact and if the strong energy condition holds, then the theorem says that there must be a singularity in the sense that the world began for some freely falling particle at a finite time or for some photon at a finite affine parameter. It does *not* say that this is necessarily a curvature singularity or a state of infinite energy density. The strong energy condition might be considered now as more questionable than in 1967. Firstly, we now believe that there is “dark energy” which violates the strong energy condition. Secondly, most inflationary scenarios violate the strong energy condition. However, the (hypothetical) idea of inflation applies to the very early universe where it is questionable if our classical (i.e., non-quantum) spacetime model is valid. So even if one accepts the idea of inflation one might argue that Theorem 1 predicts a singularity in the regime where the model of a classical spacetime is applicable.

As we don’t know if our universe is spatially compact, we would like to have another theorem which includes non-compact spatial topologies. This requires a more sophisticated formulation of the third condition, because then we do not have a spacelike compact submanifold from which the integral curves of our vector field V could start. Hawking’s second theorem reads as follows.

Theorem 2 (Hawking, 1967): Spacetime cannot be timelike and lightlike geodesically complete in the past if the following three assumptions hold:

- (a) The strong energy condition is satisfied,

$$R_{\mu\nu}K^\mu K^\nu \geq 0 \quad \text{if} \quad g_{\mu\nu}K^\mu K^\nu \leq 0.$$

- (b) The strong causality condition is satisfied, i.e., every neighbourhood of a point p contains a neighbourhood that no timelike or lightlike curve through p intersects more than once.
- (c) Let V be the past-oriented vector field whose integral curves are the timelike geodesics issuing from a point p and let $\theta = \nabla_\mu V^\mu$. Then there is a past-oriented timelike vector $w^\mu \partial_\mu|_p$ at p and a positive constant b such that on each past-oriented timelike geodesic from p the inequality $\theta < -3k/b$ holds within a proper time distance b from p , where the positive number k is defined by $k = -V^\mu|_p w_\mu$.

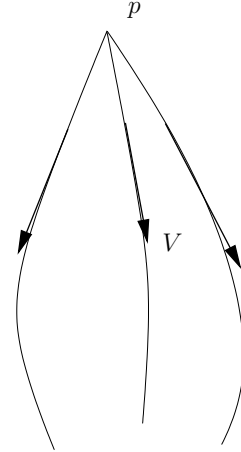


Figure 48: Illustration of Theorem 2, assumption (c)

In contrast to Theorem 1, in Theorem 2 condition (c) is now not quite so easily connected with observations. In essence, however, it just says that the expansion in the past of some event must be negative and bounded away from zero by a certain amount.

The fact that solutions to Einstein’s field equation with a reasonable matter content have a strong tendency to form singularities is viewed by many as the most serious problem of general relativity as a classical theory. It is widely believed that we will really understand what is going on near a singularity only if we have some quantum version of general relativity.